


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

Dalang Yu[†], Xiao Yang[†], Bixia Tang[†], Yi-Hsuan Pan[†], Jianing Yang[†], Guangya Duan[†], Junwei Zhu, Zi-Qian Hao, Hailong Mu, Long Dai, Wangjie Hu, Mochen Zhang, Ying Cui, Tong Jin, Cui-Ping Li, Lina Ma, Language translation team, Xiao Su, Guoqing Zhang,

Wenming Zhao and Haipeng Li 

Corresponding authors: Guoqing Zhang, E-mail: gqzhang@picb.ac.cn; Wenming Zhao, E-mail: zhaowm@big.ac.cn; Haipeng Li, E-mail: lihaipeng@picb.ac.cn; Tel: +86-21-54920460

[†]Joint authors.

Abstract

Genomic epidemiology is important to study the COVID-19 pandemic, and more than two million severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomic sequences were deposited into public databases. However, the exponential increase of sequences invokes unprecedented bioinformatic challenges. Here, we present the Coronavirus GenBrowser (CGB) based on a highly efficient analysis framework and a node-picking rendering strategy. In total, 1,002,739 high-quality genomic sequences with the transmission-related metadata were analyzed and visualized. The size of the core data file is only 12.20 MB, highly efficient for clean data sharing. Quick visualization modules and rich interactive operations are provided to explore the annotated SARS-CoV-2 evolutionary tree. CGB binary nomenclature is proposed to name each internal lineage. The pre-analyzed data can be filtered out according to the user-defined criteria to explore the transmission of SARS-CoV-2. Different evolutionary analyses can also be easily performed, such as the detection of accelerated evolution and ongoing positive selection. Moreover, the 75 genomic spots conserved in SARS-CoV-2 but non-conserved in other coronaviruses were identified, which may indicate the functional elements specifically important for SARS-CoV-2. The CGB was written in Java and JavaScript. It not only enables users who have no programming skills to analyze millions of genomic sequences, but also offers a panoramic vision of the transmission and evolution of SARS-CoV-2.

Keywords: coronavirus GenBrowser, SARS-CoV-2, genomic epidemiology, transmission, evolution

Bixia Tang is an assistant professor at Beijing Institute of Genomics. Her current research focuses on big data integration, visualization and tools integrated on cloud computing platform.

Dalang Yu is a PhD student at Shanghai Institute of Nutrition and Health. His research focuses on biological software development and genomic sequence analysis.

Xiao Yang was a computer programmer at Shanghai Institute of Nutrition and Health. He is good at software architecture, Java and JavaScript programming.

Yi-Hsuan Pan is an associate professor at East China Normal University. Her research focuses on evolutionary biology and neuroscience.

Jianing Yang is a PhD student at Shanghai Institute of Nutrition and Health. Her current research interests include population genetics and molecular evolution of viruses.

Guangya Duan is a master student at Beijing Institute of Genomics. Her current research interests include biological software's visual interface development and data processing.

Junwei Zhu is an engineer at Beijing Institute of Genomics. His research focuses on software and database system design and development.

Zi-Qian Hao is a PhD student at Shanghai Institute of Nutrition and Health. His current research interests include evolutionary biology and population genetics.

Hailong Mu was a computer programmer at Shanghai Institute of Nutrition and Health. He is good at Java program development.

Long Dai is a computer programmer at Shanghai Institute of Nutrition and Health. He is good at software architecture, Java and JavaScript programming.

Wangjie Hu is a PhD student at Shanghai Institute of Nutrition and Health. His work includes the development and application of novel demographic history inference method.

Mochen Zhang is a PhD student at Beijing Institute of Genomics. Her current research interest is biological software development.

Ying Cui is a master student at Beijing Institute of Genomics. Her current research interest is algorithmic research on image recognition.

Tong Jin is a master student at Beijing Institute of Genomics. She is good at analyzing multi-omics data with bioinformatics method.

Cui-Ping Li is an assistant professor at Beijing Institute of Genomics. Her current research interests include genomic variation analysis and population genetic analysis.

Lina Ma is an associate professor at Beijing Institute of Genomics. Her current research interests include big data integration and analytics, and multi-omics annotation of non-coding RNAs. Language translation team is composed of teachers and students at Beijing Language and Culture University.

Language translation team is a group at Beijing Language and Culture University: Ning He, Ting Zhou, Huang Li, Yuhong Guan, Qin Zhou, Yuyuan Guo, Jing Lv, Ting Peng, Nan Yang, Siyi Hou, Jingxuan Yan, Chenglin Zhu, Wenjing Liu, Huanxiao Song, Han Gao, Jinglan He, Tiantian Li, Ruiwen Fei, Shumei Zhang.

Xiao Su is a Professor at Institut Pasteur of Shanghai. He focuses on respiratory infection and histopathology.

Guoqing Zhang is a professor at Biomedical Big Data at Shanghai Institute of Nutrition and Health. He is the vice director at Biomedical Data Center and focuses on developing new online database and service to help the application of precise medicine, human phonemics and microbiome.

Wenming Zhao is a professor at Beijing Institute of Genomics. He focuses on the methodology research for next generation sequencing data analysis and the bioinformatics database construction. His team constructed the Genome Sequence Archive database in China, which promotes the scientific data centralized management and sharing.

Haipeng Li is a professor at Evolutionary Genomics at Shanghai Institute of Nutrition and Health. His group focuses on developing new evolutionary theory to understand key biological questions and make novel findings.

Received: August 9, 2021. **Revised:** November 26, 2021. **Accepted:** December 20, 2021

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

Real-time tracking of the transmission and evolution of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is essential for public health during the COVID-19 pandemic [1, 2]. Since January 2020 more than two million genomic sequences have been deposited into public databases, such as National Center for Biotechnology Information (NCBI) GenBank [3], Global Initiative on Sharing All Influenza Data (GISAID) [4, 5]. The exponential increase of genomic sequences provides a great opportunity to monitor the transmission and evolution of SARS-CoV-2 but invokes unprecedented bioinformatic challenges.

Several web browsers have been developed to analyze the genomic data and track the COVID-19 pandemic. The UCSC SARS-CoV-2 Genome Browser was derived from the well-established UCSC Genome Browser for visualization of nucleotide and protein sequences, sequence conservations and many other properties of wild-type and variants of SARS-CoV-2 [6]. The WashU Virus Genome Browser provides Nextstrain-based phylogenetic-tree view and genomic-coordinate, track-based view of genomic features of viruses [7]. The pathogen genomics platform Nextstrain allows analysis of genomic sequences of approximately 4000 strains of SARS-CoV-2 and investigation of its evolution [8], which cannot timely analyze millions of increasing genomic sequences. Therefore, new approaches are essential and indispensable to enable users easier to explore the large amount of SARS-CoV-2 genomic sequences.

In this study, we developed the Coronavirus GenBrowser (CGB) (Figure 1). All the high-quality genomic sequences and the associated transmission-related metadata were timely integrated and analyzed to provide the latest panoramic view of the pandemic. To investigate a local transmission, the data can be easily filtered according to countries, regions, keywords and the collection date of viral strains. Thus, even if users do not have any programming skills, the CGB enables them to efficiently explore millions of SARS-CoV-2 genomic sequences and monitor the global/local transmission and evolution of SARS-CoV-2. All the pre-analyzed genomic mutations and the associated metadata can be easily downloaded, reanalyzed and reshared. Since a cleaned genome alignment with almost no ambiguous nucleotide sites can be easily reconstructed from the CGB core data file, the CGB may provide a great convenience for the society to study viral evolution further.

Material and methods

Data quality control and distributed genome alignments

SARS-CoV-2 genomic variations were obtained from the 2019nCoV database [9] established by China National Center for Bioinformation (CNCB) [10], as an integrated

resource based on Global Initiative on Sharing All Influenza Data (GISAID) [4, 5], NCBI GenBank [3], China National GeneBank DataBase (CNGDB) [11], the Genome Warehouse (GWH) [12] and the National Microbiology Data Center (NMDC, <https://nmdc.cn/>). Detailed information on this database is available at https://ngdc.cncb.ac.cn/ncov/release_genome. All SARS-CoV-2 strains were isolated from humans, and quality control was applied to obtain high-quality SARS-CoV-2 genomic sequences (Figures S1 and S2, see Supplemental Materials and Methods Section 1). Because of the explosion in SARS-CoV-2 genomic data, the distributed alignment system was developed to enable daily update (Figure 1), which reduces the total alignment time complexity to $\mathcal{O}(n)$, where $\mathcal{O}(\cdot)$ is a linear function, and n is the number of viral strains (see Supplemental Materials and Methods Section 2).

Reconstruction and timely update of the annotated phylogenetic tree

Before October 2020 all high-quality sequences in distributed alignments were analyzed as a whole and used to reconstruct the evolutionary tree (see Supplemental Materials and Methods Section 4). After October 2020, the efficiency of this process became too low to perform timely updates. Therefore, new trees were reconstructed by appending new sequences to existed tree (see Supplemental Materials and Methods Section 8). Ambiguous and missing nucleotides were imputed by incorporation of the neighboring lineages (Figure S3, see Supplemental Materials and Methods Section 5) and mutations in strains of each branch are recapitulated according to the principle of parsimony [13, 14] (see Supplemental Materials and Methods Section 6). A highly effective maximum-likelihood method (TreeTime) is used to determine the dates of internal nodes [15], as it allows fast inference by ‘the post- and pre-order traversals’ with tabulated key values for back tracing. This algorithm was implemented in the CGB with very minor revisions (Figures S4 and S5; see Supplemental Materials and Methods Section 7). The genome-wide mutation rate was also timely calculated (see Supplemental Materials and Methods Section 13).

Displaying SARS-CoV-2 genomic mutations in tree-based format

Similar to NextStrain [8] and the WashU Virus Genome Browser [7], the CGB uses a tree-based file format to store SARS-CoV-2 genomic mutations. The head of the core data file contains the data version, the updated date, the genomic region analyzed and the mutation rate estimated for each gene. The core file also contains information on collection date, gender and age of patient, location for each strain, mutations and inferred date for each internal node. To allow fast access to the data, redundant information has been minimized. The detailed description of data format is available in the online readme (<https://ngdc.cncb.ac.cn/ncov/apis/>).

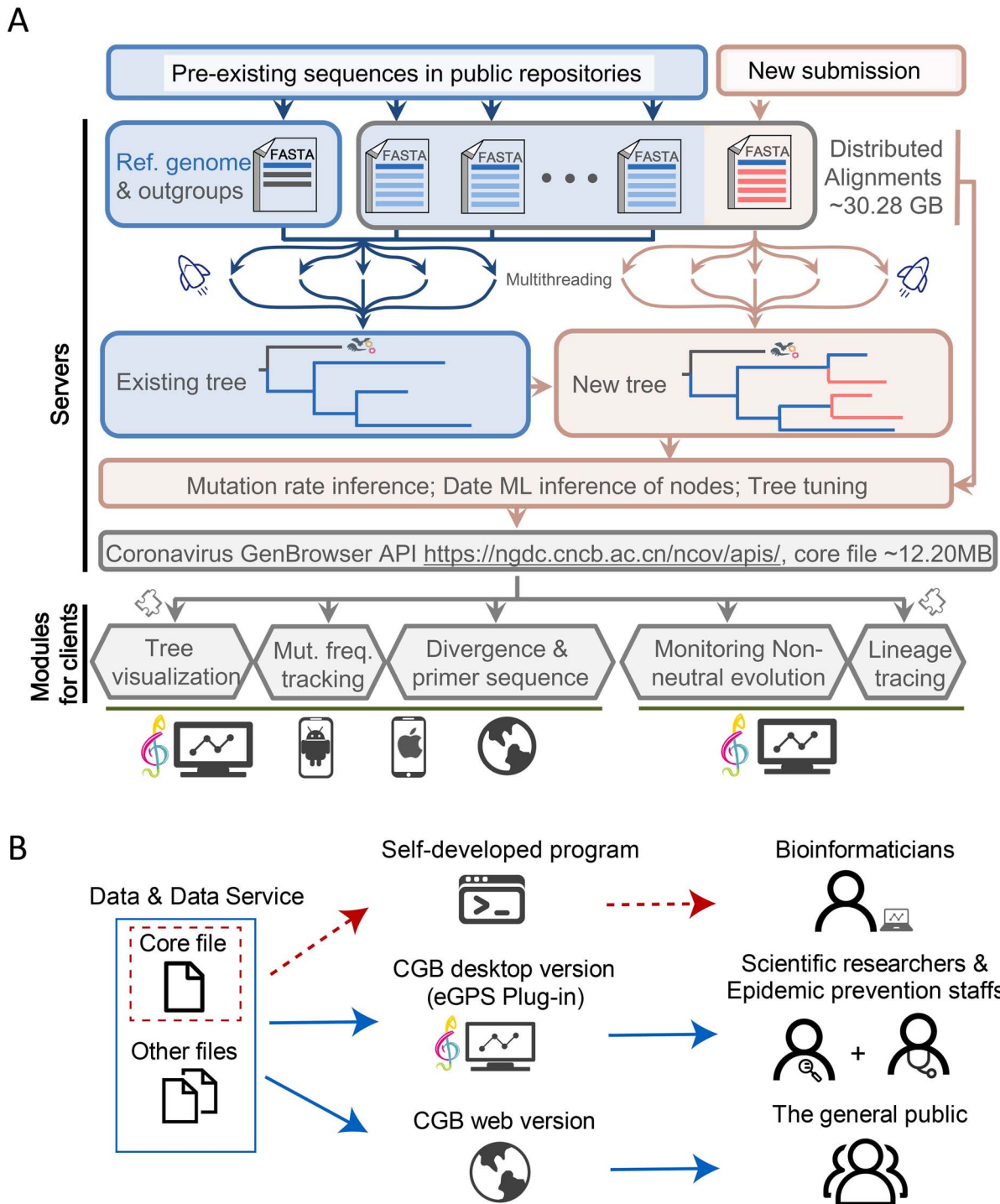


Figure 1. Coronavirus GenBrowser. (A) Timely updates of SARS-CoV-2 genomic data and visualization framework of CGB. The core file includes the pre-analyzed genomic mutations of SARS-CoV-2 and the associated metadata. All timely updated data can be freely accessed at <https://ngdc.cncb.ac.cn/ncov/apis/>. (B) Illustration of three hypothetical user groups of the CGB.

Visualization of the huge evolutionary tree by node-picking rendering strategy

When visualizing the huge evolutionary tree, the most of lineages are invisible because lineages overlap each other. Thus, a node-picking rendering strategy was implemented to efficiently visualize huge evolutionary tree, in which only surface lines will be painted. If the tree is zoomed in, only the visible sub-area of the tree will

be painted. Using this strategy, millions of lineages can be visualized effectively (Figure S6, see Supplemental Materials and Methods Section 9).

CGB binary nomenclature and data searching

A number of different naming systems have been proposed [16, 17], but these systems only name a few internal nodes or branches. As there are a large number of

internal nodes on the huge evolutionary tree, the CGB binary nomenclature was developed following the most recent common ancestor (MRCA) concept (Figure S12) to obtain CGB ID for each node (see Supplemental Materials and Methods Section 12). CGB ID can be used to search a specific lineage. Moreover, isolate names, accession numbers and mutations are also searchable (see Supplemental Materials and Methods Section 11).

Mutation analysis

A root-to-tip linear regression method [18] was used to estimate the mutation rate of SARS-CoV-2. For each strain with a different collection date in a tip-dated time tree, the number of mutations, including that of recurrent mutations, was counted subsequent to the appearance of MRCA (see Supplemental Materials and Methods Section 13). To avoid the effect of recombination, recombination flag is labeled for each mutation by analyzing hybrid genomic structure (Figure S13, see Supplemental Materials and Methods Section 14).

Lineage tracing

For lineage tracing, genomic sequences of SARS-CoV-2 strains collected from patients or environments are used as the queries. These query sequences should be aligned with the reference genomic sequence of SARS-CoV-2 (GenBank accession number: NC_045512) [19]. A very fast algorithm was implemented to count the difference between a query sequence and the genomic sequence of a node. For one query, nodes with the least difference are considered as its candidate targets.

Detection of branch-specific accelerated evolution of SARS-CoV-2

To detect branch-specific accelerated evolution, each internal branch of the SARS-CoV-2 tree was examined. For each internal branch, the observed number of mutations of the i th gene ($\gamma_{obs,i}$) was compared with the expected number of mutations of the same gene ($\gamma_{exp,i}$). The significance level of acceleration was determined by Poisson probability [20, 21]. It is a one-tailed test. The condition $t > 10$ (days) was used for detection of branch-specific accelerated evolution of SARS-CoV-2.

Detection of ongoing selection of SARS-CoV-2

To detect ongoing positive selection, allele frequency trajectory with an S-shaped curve was examined (Table S3, Figures S14 and S15). To reduce the impact of hitchhiking by neutral mutation, only non-synonymous mutations were analyzed, although non-coding mutations [22] can also be beneficial (see Supplemental Materials and Methods Section 16).

Local analysis for new SARS-CoV-2 genomic data

The de novo tree building and tree appending program was implemented in the eGPS software (http://www.egps-software.net/egpscloud/eGPS_Desktop.html) [23]. Users can download a compressed CGB data file

(<https://ngdc.cncb.ac.cn/ncov/apis/>) and analyzed together with their own SARS-CoV-2 genomic data. This function ensures that a timely analysis can be easily performed. Technical details were provided in the Supplemental Materials and Methods. The detailed description how to use it is provided in the user manual of the eGPS software.

Data source

For genomic sequence alignments, high-quality SARS-CoV-2 genomic variations were obtained from the 2019nCoV database [9, 24], which is an integrated resource based on GenBank, GISAID [4, 5], CNGBdb [11], the GWH [12] and the NMDC (<https://nmdec.cn/>). The data version 'data.2021-06-07' was used for the analysis ($n = 1,002,739$) in this paper and it can be downloaded freely from <https://ngdc.cncb.ac.cn/ncov/apis/archives/>. The current data contain more than two million SARS-CoV-2 genomic sequences when this manuscript was revised.

Results

The construction of a million-level evolutionary tree

After quality control, 1,002,739 high-quality genomic sequences were obtained for subsequent analyses. The number of identified high- and low-quality genomes in each month is summarized (Figure S2). To allow timely analysis of a large number of sequences, we first solved the problem that all viral genomic sequences have to be realigned when nucleotide sequences of new genomes become available. This is extremely time-consuming. With the distributed alignment system (Figure 1A), we dramatically reduced the total time required for the alignment. We also built the evolutionary tree on the existed tree with new genomic data in order to reduce the complexity of tree construction. With these modifications, a tremendous evolutionary tree can be reconstructed for each update, millions of SARS-CoV-2 genomic sequences can be timely analyzed, and data can be easily shared, reanalyzed and reconstructed (Figure 1A).

For the huge evolutionary tree, mutations on each branch were identified according to the principle of parsimony [13, 14] and the dates of internal nodes were inferred with minor revisions of a highly effective maximum-likelihood method (TreeTime) [15]. The pre-analyzed genomic mutations of SARS-CoV-2 and the associated metadata are shared to the general public in a tree-based CGB format. The size of distributed alignments is 30.28 GB for the 1,002,739 SARS-CoV-2 genomic sequences. The tree-based data format allows the compression ratio to reach 2541:1, meaning that the size of compressed core file containing the tip-dated evolutionary tree, the pre-analyzed genomic mutations and associated metadata is as small as 12.20 MB with zip compression (Figure 1A). Whenever necessary, by using the core file and the SARS-CoV-2

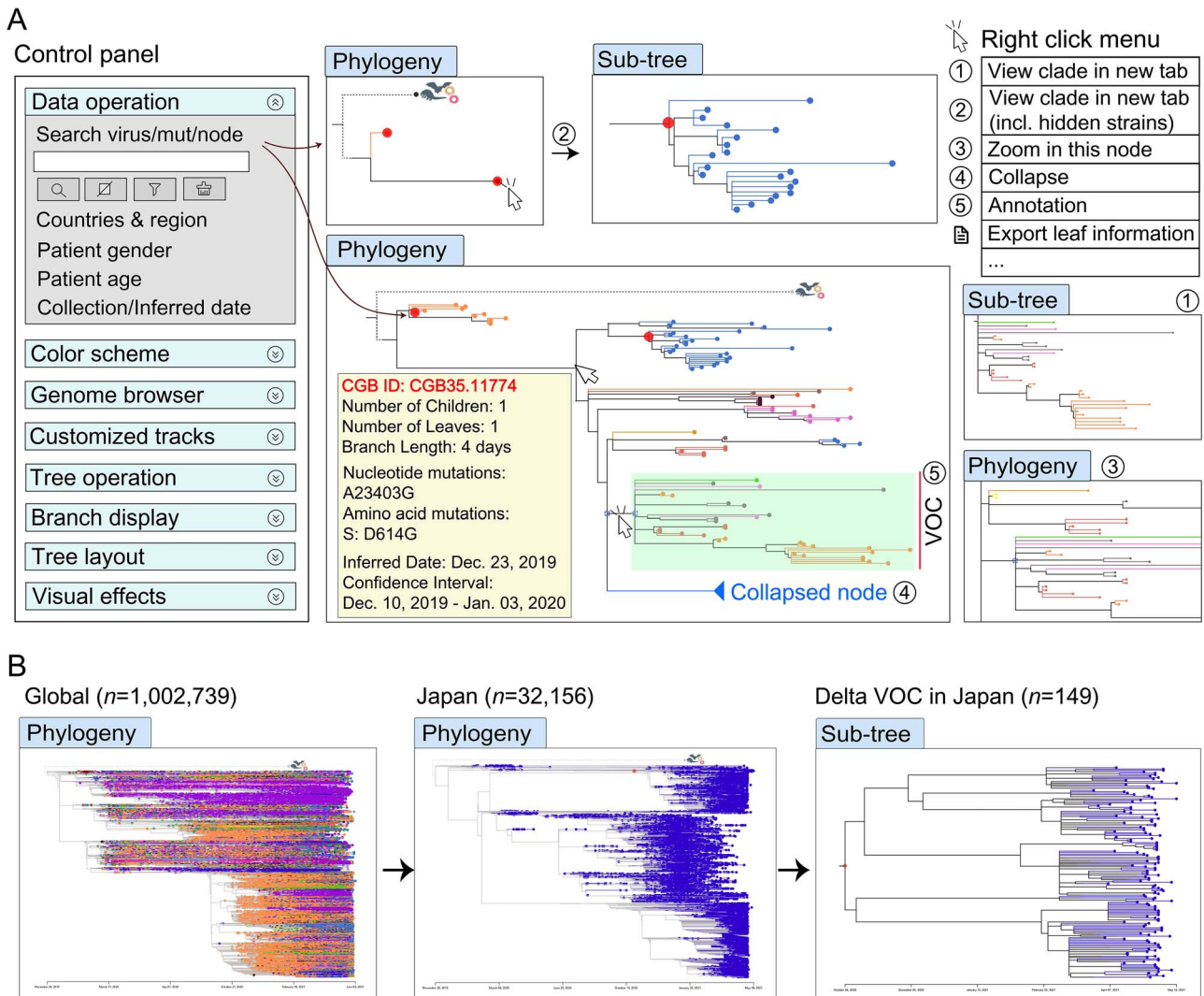


Figure 2. An illustration of CGB interactive functionalities. **(A)** A sketch of the tree visualization panel to present the major CGB functions. Strains, mutations and CGB IDs can be searched. The searched nodes are highlighted and can be visualized individually. A number of frequently used menus are presented at the top right corner. **(B)** An example for how to view the transmission and evolution of the Delta VOC in Japan. First, users can filter for strains collected from Japan. Search the CGB ID of the Delta VOC (CGB531065.580055) and then right click the highlighted node and choose the menu item of View clade in new tab in the pop-up menu.

reference genomic sequence, cleaned genomic sequence alignment with almost no ambiguous nucleotide sites can be reconstructed for all considered viral isolates (extra programming skills are needed) (Figure 1B). Thus, this volume-saving core file format ensures low-latency access to the data and enables fast sharing and reanalysis of a large number of SARS-CoV-2 genomic variants. The latest core file (named as mainDataFile) can be downloaded from <https://ngdc.cncb.ac.cn/ncov/apis/data-latest/>.

Highly efficient tree visualization and operation

To efficiently visualize the results, a node-picking rendering strategy was implemented for painting the evolutionary tree, indicating that only parts of the tree are painted (Figure S6, see Supplemental Materials and Methods Section 9). In another word, only elements shown on the screen and visible to the user are

painted. This design makes the visualization process highly efficient, and the evolutionary tree of more than one million strains can be visualized. It takes about one second for the visualization process in different operation systems (Table S1).

The CGB is also a highly efficient platform to search or filter variants based on transmission-related meta-data (Figure 2). Useful interactive functionalities were developed to navigate users through the huge tip-dated evolutionary tree. First, users can easily search internal branches or variants with certain mutations, or isolate names of virus. There are 400,298 internal branches in the evolutionary tree ($n = 1,002,739$), and each branch has been named by CGB binary nomenclature (i.e. CGB ID) (Figure S12) and is searchable. Thus, different variants of concerns (VOCs) can be easily identified and visualized on the huge evolutionary tree (Table 1). Second, users can easily filter out the data according

Table 1. CGB ID for VOC

WHO label	Pango lineage	Documented samples	CGB ID	Defining SNPs ^a
Alpha	B.1.1.7	UK, September 2020	CGB84017.91425	ORF1ab: T1001I, A1708D, I2230T, SGF3675-S: HV69-, Y144-, N501Y, A570D, P681H, T716I, S982A, D1118H ORF8: Q27*, R52I, Y73C N: D3L, S235F
Beta	B.1.351	South Africa, May 2020	CGB391494.393307	E: P71L N: T205I ORF1a: K1655N S: D80A, D215G, K417N, A701V, N501Y, E484K
Gamma	P.1	Brazil, November 2020	CGB222196.451180	ORF1ab: S1188L, K1795Q, SGF3675-S: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I ORF3a: G174C ORF: E92K N: P80R
Delta	B.1.617.2	India, October 2020	CGB531065.580055	S: T19R, L452R, T478K, P681R, D950N ORF3a: S26L M: I82T ORF7a: V82A, T120I N: D63G, R203M, D377Y

^aThe information of defining SNPs was obtained from the Pango lineages [31].

to the collection date, the country/region and the gender and age groups. Third, the visualization of a sub-clade in another tab is allowed. Forth, different annotations are provided to mark the clades of interest. All those features have brought great convenience for users.

Coordinated annotation tracks

Six user-friendly tracks are visualized with rapid response time. Coordinated annotation tracks are provided to show genome structure and key domains, allele frequencies, sequence similarity between various coronavirus, multi-genome alignment and primer sets for detection of SARS-CoV-2 (Figures S8–S11, see Supplemental Materials and Methods Section 10). The track-related data files are available via <https://ngdc.cncb.ac.cn/ncov/apis/data-latest/>.

Transmission case studies through CGB

Based on the large data volume and user-friendly CGB, many analyses can be quickly conducted. Take the spread of VOCs in India as an example. Among the 1,002,739 strains, there were 3349 ones sampled in India. Nearly all major SARS-CoV-2 lineages can be found in India in different stages of the pandemic (Figure 3). By searching the CGB IDs of the VOCs (Table 1), the clades of VOCs were identified. In total, there were 464 Delta (B.1.617.2) strains (464/3349=13.85%), 185 Alpha (B.1.1.7) strains (185/3349=5.52%) and 11 Beta strains (11/3349=0.32%) in the Indian sample. The ratios change when only considering recent viral strains collected after 1 April 2021. There were 589 Indian samples after date filtering. Among them, there were 364 Delta strains (364/589=61.80%), 55 Alpha strains (55/589=9.34%) and one Beta strains (1/589=0.17%). Thus, the Delta variant increased more rapidly than others, and the most of recent infections in India are caused by the Delta VOC.

Another three examples were provided to show that the CGB is an efficient platform to investigate local and global transmission of COVID-19 (Figure 4). To trace the

origin of a local COVID-19 outbreak, the lineage tracing was implemented in the CGB. The closest nodes were revealed for the three outbreaks in China during this year, which indicates different origin of the three outbreaks. Their neighboring strains can be viewed individually and further investigated. The analysis is extremely fast and can be performed on a desktop computer. Therefore, the CGB is a highly efficient platform to investigate the origin of a local COVID-19 outbreak.

Mutation analysis

The CGB also estimates the mutation rate of whole genome and each gene (Table S2). Applying 1,002,739 genomic sequences, the estimated genome-wide mutation rate is 1.0794×10^{-3} per nucleotide per year. The mutation rate is variable for different genes (Table S2). The latest estimates of mutation rate are provided in the core data file (named as mainDataFile) that can be downloaded from <https://ngdc.cncb.ac.cn/ncov/apis/data-latest/>.

We also found that the mutation rate could be different among sites. Using the CGB core data file, we conducted a 10-base sliding window analysis with a sliding step of one base and identified fine-scaled mutation cold spots along the viral genome (see Supplemental Materials and Methods Section 15), indicating the genomic regions with mutation rate significantly lower than the average mutation rate of the entire genome. In total, 657,074 (recurrent) mutations were identified and 868 mutation cold spots were found with a false discovery rate (FDR) corrected P -value < 0.01 (Figure 5, Supplemental excel file). The coldest spot is located in ORF1ab, which encodes nsp13 helicase (nucleotides 16,294–16,307) (FDR corrected P -value = 4.79×10^{-46}). Interestingly, it has been found that sequence conservation is restricted to ORF1ab:nsp10–13 among 14 coronaviruses [25, 26]. It indicates that nsp13 helicase might be essential for coronaviruses and SARS-CoV-2. Moreover, among the 868 mutation cold spots, there are 75 conserved spots in SARS-CoV-2, but not conserved among other

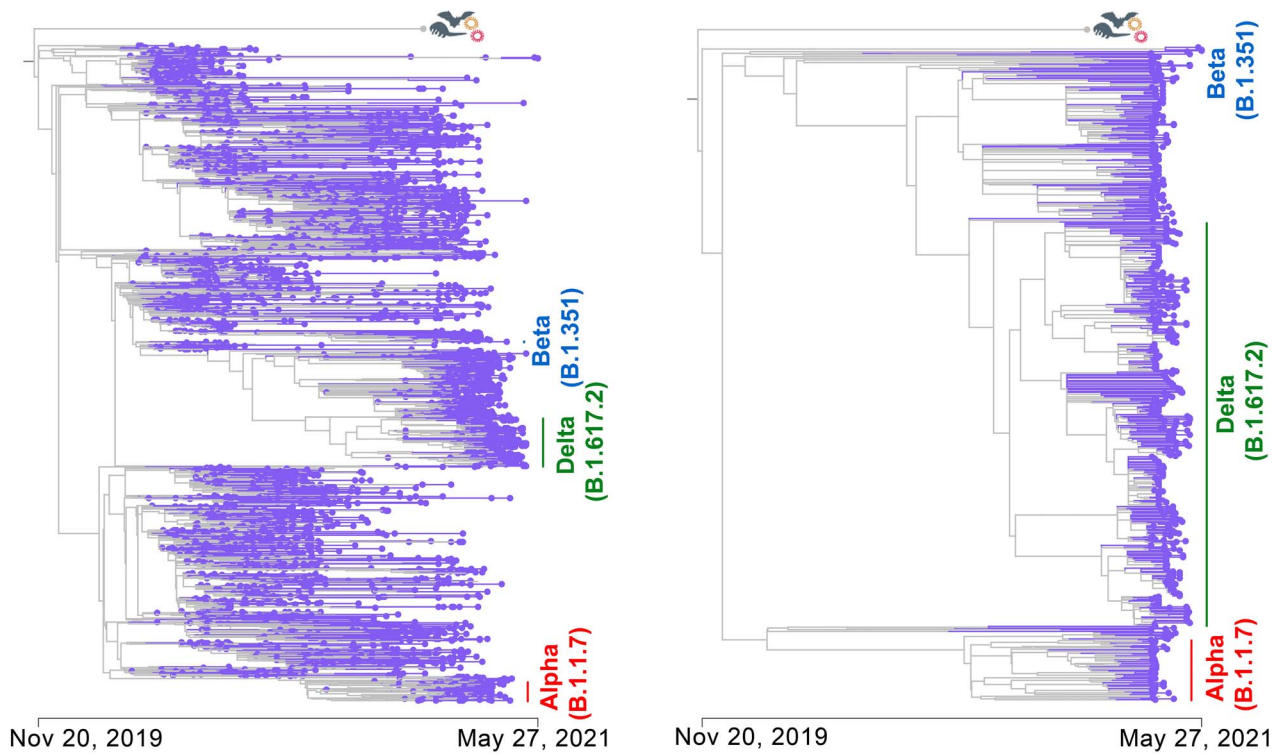


Figure 3. Transmission-related study using the CGB. Proportion change of VOCs in India. The tree on the left shows all Indian samples of 1,002,739 SARS-CoV-2 strains and the tree on the right shows Indian samples after 1 April 2021. Different VOCs are annotated in different colors, the Alpha (B.1.1.7) VOC in red, the Delta (B.1.617.2) VOC in green and Beta (B.1.351) VOC in blue. The Gamma VOC is not found in India.

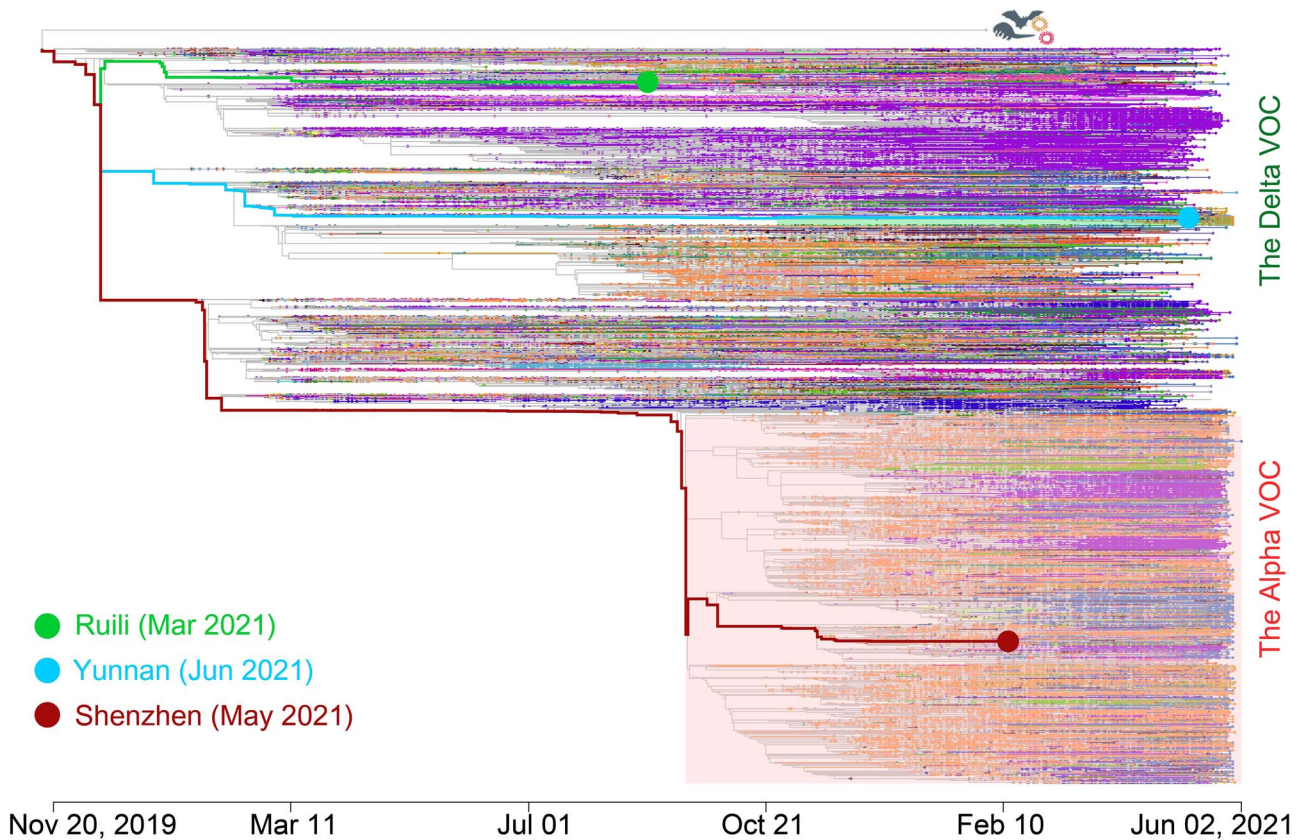


Figure 4. Tracing the origin of three local outbreak of COVID-19 in China. One sequence sampled during each outbreak is used as the query sequence which is not included in the CGB dataset version used in analysis. Their closest targets were marked with colored dot and the evolutionary paths were highlighted. The GISAID IDs for the queries are EPI_ISL_1595852 (Ruili), EPI_ISL_2834004 (Yunnan) and EPI_ISL_2405168 (Shenzhen).

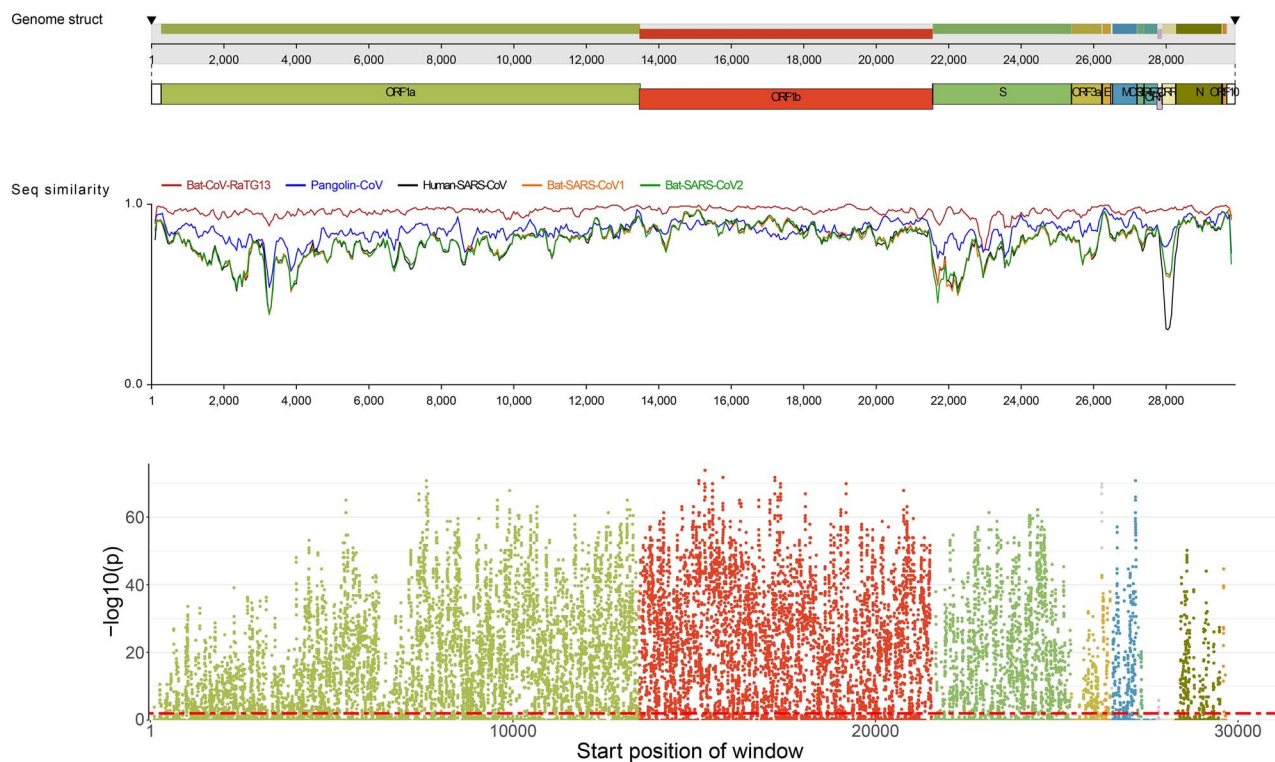


Figure 5. Manhattan plot of mutation cold spots in the genome of SARS-CoV-2. Results of genome-wide scan for mutation cold spots are shown in Manhattan plot of significance against SARS-CoV-2 reference genomic locations. In total, 1,002,739 high-quality genomic sequences were analyzed. Each dot represents one window. P -values are FDR-corrected. The dotted red line denotes FDR-corrected P -value < 0.01 . Dots above the line represent mutation cold spots. Genomic structure and sequence similarity between SARS-CoV-2 reference genome (NC_045512.2) [19] and the genomes of five other coronaviruses are shown above the Manhattan plot.

coronaviruses. These SARS-CoV-2-specific conserved elements may play key roles for SARS-CoV-2-specific functions.

The detection of accelerated evolution

The CGB provides a module to detect branch-specific accelerated evolution of SARS-CoV-2. We found that within 186,746 internal branches with $t > 10$ (days) on the evolutionary tree ($n = 1,002,739$), 70 branches were detected to have a genome-wide accelerated evolution (FDR corrected $P < 0.05$), 332 branches were detected to have an accelerated evolution of ORF1ab (FDR corrected $P < 0.05$), and two branches was found to have an accelerated evolution of the spike gene (FDR corrected $P < 0.05$) (Supplemental excel file). These evolution-accelerated variants can be used for future studies.

The detection of ongoing positive selection

The mutation frequency trajectory for each mutation can be easily visualized by using the CGB. It has a module to detect ongoing positive selection based on S-shaped frequency trajectory of a selected allele (Figures S14 and S15, see Supplemental Materials and Methods Section 16). It has been shown that the SARS-CoV-2 variant with G614 spike protein has a fitness advantage [27]. Our analysis using the CGB confirmed this finding even when the

G614 frequency was very low ($< 10\%$) (Figure S16), indicating that the CGB can detect putative advantageous variants before they become widely spread. The CGB also predicted an increase in the frequency of S:p.P681R of the Delta VOC (Figure S16), suggesting that variants with the mutation may be advantageous. S:p.P681R is located on the spike S1/S2 cleavage site, and another mutation (S:p.P681H) on the same position has been found to be advantageous [28] and may interact with other mutations [22] in the Alpha VOC. Based on 1,002,739 samples, the CGB detected 13 putative advantageous mutations in the spike protein (Table S4). However, as an increase in mutation frequency could be due to sampling bias and epidemiological factors [27], putative advantageous variants should be closely monitored.

Conclusion

In this study, we developed an effective surveillance tool for the transmission and evolution of SARS-CoV-2. It can process million-level genomic sequences and the associated transmission-related metadata (Figure 1A). A highly efficient visualization module is established, and rich interactive operations are allowed to explore the annotated evolutionary tree (Figure 2). We investigated three local COVID-19 outbreaks by the searching and lineage tracing functionalities (Figure 4). We also implemented a new method to detect ongoing positive

selection for each viral non-synonymous mutation. The branches with accelerated or reduced evolutionary rate are identified to provide a real-time tracking on the change of evolutionary rate, which could reveal epidemic factors affecting the viral transmission.

The CGB also provides an efficient way for clean data sharing (Figure 1B). It could be difficult for researchers to download all the raw genomic sequences, perform data quality control and analyze the large amount of data by their own. By examining literatures related to this topic [29, 30], most of those studies have similar methods to prepare alignments, build phylogenetic tree, infer mutations and date each internal node. The data preparation procedure is tedious but may require some skills and is often time consuming when the sample size is extremely large. Nevertheless, by downloading the pre-analyzed CGB data files, users can have all those clean data in a few minutes.

During the analysis, we noticed that a very small percentage of sequences have abnormal collection dates, which could severely skew the evolutionary tree. After examining all possible reasons, it is likely due to that the year of collection date was incorrectly filled in the most cases. We then deleted these sequences although their sequence quality is high. Thus, we would suggest that researchers could pay attention on the year of collection date when submitting their sequences.

When millions of SARS-CoV-2 high-quality genomic sequences are available, interactive visualization of SARS-CoV-2 evolutionary tree is essential for effectively exploring the viral evolution and transmission. However, the Nextstrain-based phylogenetic-tree view only supports about 4000 strains of SARS-CoV-2 [7, 8]. Therefore, the CGB not only enables users to timely analyze millions of viral genomic sequences, but also implements a number of down-stream evolutionary analysis (Figures 1 and 2). Moreover, a volume-saving data format is used to store the information of tree, mutations and metadata. With the core file CGB provided and the SARS-CoV-2 reference genomic sequence, cleaned genomic sequence alignment can be reconstructed for further analysis without downloading all the sequences, performing quality controls and doing sequence alignment again. Therefore, it is believed that the CGB will accelerate many researches in this field.

The public science education is extremely important for the anti-epidemic to show that SARS-CoV-2 has been evolving. Therefore, a web-based CGB was also developed. It is a simplified version of CGB that provides a convenient way to access the data via a web browser, such as Google Chrome, Firefox and Safari (Figure S7). For educational purpose, nine language versions (Chinese, English, German, Japanese, French, Italian, Portuguese, Russian and Spanish) are available. The web-based CGB package can be downloaded and reinstalled on any websites. Two pre-installed websites are provided (<https://www.biosino.org/cgb/> and <https://ngdc.cncb.ac.cn/cgb/>). Moreover, to accelerate the access, a sub-tree was used that keeps the

topological structure of the original evolutionary tree and the CGB IDs of major branches unchanged (see Supplemental Materials and Methods Section 17).

Users are recommended to download the desktop version to access the full functions of CGB. The desktop version is provided as a plug-in module for the eGPS software (http://www.egps-software.net/egpscloud/eGPS_Desktop.html). To promote a timely analysis of newly sequenced genomic data of SARS-CoV-2, users can perform a local analysis to analyze these new data, together with the public data globally sampled. Overall, the CGB is frequently updated which provides a timely panoramic vision of the global and local transmission and evolution of SARS-CoV-2.

Key Points

- A highly efficient analysis framework was developed and integrated in the CGB to timely analyze millions of SARS-CoV-2 genomic sequences.
- Visualization modules were developed to enable users to interactively explore the tremendous annotated SARS-CoV-2 evolutionary tree.
- Genomic spots conserved in SARS-CoV-2 but non-conserved in other coronaviruses may indicate the functional elements specifically important for SARS-CoV-2.
- A timely updated panoramic vision of the transmission and evolution of SARS-CoV-2 was provided.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Data availability

All timely updated data are freely available at <https://ngdc.cncb.ac.cn/ncov/apis/>. The desktop standalone version (Figure S7A) provides the full function of CGB and has a plug-in module for the eGPS software (http://www.egps-software.net/egpscloud/eGPS_Desktop.html) [23]. We provide friendly documentation on how to run the software/.

Members of the language translation team

German: Ning He, Jing Lv, Ting Peng.

Italian: Ting Zhou, Nan Yang, Siyi Hou.

Portuguese: Huang Li, Jingxuan Yan, Chenglin Zhu, Wenjing Liu.

Russian: Yuhong Guan, Huanxiao Song.

Spanish: Qin Zhou, Han Gao, Jinglan He, Tiantian Li, Ruiwen Fei, Shumei Zhang.

French: Yuyuan Guo.

Acknowledgments

We thank Ya-Ping Zhang for providing valuable advice 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 grants from the National Key Research and Development Project of China (No. 2020YFC084-7000), from the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB38030100), Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01) and from Shanghai Institute of Nutrition and Health (No. JBGSRWBD-SINH-2021-10). Funding for open access charge: Shanghai Institute of Nutrition and Health.

References

1. Fineberg HV, Wilson ME. Epidemic science in real time. *Science* 2009;**324**:987.
2. Yu W-B, Tang G-D, Zhang L, et al. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. *Zool Res* 2020;**41**:247–57.
3. Sayers EW, Beck J, Bolton EE, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2021;**49**:D10–7.
4. Shu YL, McCauley J. GISAID: global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance* 2017;**22**:2–4.
5. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* 2017;**1**:33–46.
6. Fernandes JD, Hinrichs AS, Clawson H, et al. The UCSC SARS-CoV-2 Genome Browser. *Nat Genet* 2020;**52**:986–91.
7. Flynn JA, Purushotham D, Choudhary MNK, et al. Exploring the coronavirus pandemic with the WashU Virus Genome Browser. *Nat Genet* 2020;**52**:986–1001.
8. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018;**34**:4121–3.
9. Zhao W-M, Song S-H, Chen M-L, et al. The 2019 novel coronavirus resource. *Hereditas (Beijing)* 2020;**42**:212–21.
10. Xue YB, Bao YM, Zhang Z, et al. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic Acids Res* 2021;**49**:D18–28.
11. Chen F, You L, Yang F, et al. CNGBdb: China National GeneBank DataBase. *Hereditas (Beijing)* 2020;**42**:799–809.
12. Chen M, Ma Y, Wu S, et al. Genome Warehouse: a public repository housing genome-scale data. *Genomics Proteomics Bioinformatics* 2021.
13. Sankoff D. Minimal mutation trees of sequences. *SIAM J Appl Math* 1975;**28**:35–42.
14. Hartigan JA. Minimum mutation fits to a given tree. *Biometrics* 1973;**29**:53–65.
15. Sagulenko P, Puller V, Neher RA. TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol* 2018;**4**:vex042.
16. Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev* 2020;**7**:1012–23.
17. Forster P, Forster L, Renfrew C, et al. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* 2020;**117**:9241–3.
18. Bouckaert R, Vaughan TG, Barido-Sottani J, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 2019;**15**.
19. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature* 2020;**579**:265–9.
20. Ohta T, Kimura M. On the constancy of the evolutionary rate in cistrons. *J Mol Evol* 1971;**1**:18–25.
21. Wang Y, Dai G, Gu Z, et al. Accelerated evolution of an *Lhx2* enhancer shapes mammalian social hierarchies. *Cell Res* 2020;**30**:408–20.
22. Yang J, Zhang G, Yu D, et al. A Kozak-related non-coding deletion effectively increases B.1.1.7 transmissibility. *bioRxiv* 2021.
23. 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.
24. 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**:705–8.
25. Ruan YJ, Wei CL, Ee LA, et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* 2003;**361**:1779–85.
26. He JF, Peng GW, Min J, et al. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* 2004;**303**:1666–9.
27. 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**:812–27.
28. 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. *virological.org* 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> (January 2022, date last accessed).
29. Hodcroft EB, Zuber M, Nadeau S, et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* 2021;**595**:707–12.
30. Deng XD, Gu W, Federman S, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* 2020;**369**:582–7.