Genomics Proteomics Bioinformatics 18 (2020) 749–759

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

DATABASE

The Global Landscape of SARS-CoV-2 Genomes, Variants, and Haplotypes in 2019nCoVR

Shuhui Song [1,2,3,4,#], Lina Ma [1,2,3,#], Dong Zou [1,2,3,#], Dongmei Tian [1,2,#], Cuiping Li [1,2,#], Junwei Zhu [1,2,#], Meili Chen [1,2,3], Anke Wang [1,2], Yingke Ma [1,2], Mengwei Li [1,2,3,4], Xufei Teng [1,2,3,4], Ying Cui [1,2,3,4], Guangya Duan [1,2,3,4], Mochen Zhang [1,2,3,4], Tong Jin [1,2,3,4], Chengmin Shi [1,5], Zhenglin Du [1,2,3], Yadong Zhang [1,2,3,4], Chuandong Liu [1,5], Rujiao Li [1,2,3], Jingyao Zeng [1,2,3], Lili Hao [1,2,3], Shuai Jiang [1,2], Hua Chen [1,4,5,6], Dali Han [1,4,5], Jingfa Xiao [1,2,3,4], Zhang Zhang [1,2,3,4,*], Wenming Zhao [1,2,3,4,*], Yongbiao Xue [1,2,4,*], Yiming Bao [1,2,3,4,*]

[1] *China National Center for Bioinformation, Beijing 100101, China*
[2] *National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[3] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[4] *University of Chinese Academy of Sciences, Beijing 100049, China*
[5] *CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[6] *Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China*

Received 18 July 2020; revised 17 September 2020; accepted 24 September 2020
Available online 28 December 2020

Handled by Feng Gao

KEYWORDS

2019nCoVR;
SARS-CoV-2;
Database;
Genomic variation;
Haplotype

Abstract On January 22, 2020, China National Center for Bioinformation (CNCB) released the 2019 Novel Coronavirus Resource (**2019nCoVR**), an open-access information resource for the severe acute respiratory syndrome coronavirus 2 (**SARS-CoV-2**). 2019nCoVR features a comprehensive integration of sequence and clinical information for all publicly available SARS-CoV-2 isolates, which are manually curated with value-added annotations and quality evaluated by an automated in-house pipeline. Of particular note, 2019nCoVR offers systematic analyses to generate a dynamic landscape of SARS-CoV-2 **genomic variations** at a global scale. It provides all identified variants and their detailed statistics for each virus isolate, and congregates the quality score, functional annotation,

* Corresponding authors.
E-mail: baoym@big.ac.cn (Bao Y), ybxue@big.ac.cn (Xue Y), zhaowm@big.ac.cn (Zhao W), zhangzhang@big.ac.cn (Zhang Z).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2020.09.001

and population frequency for each variant. Spatiotemporal change for each variant can be visualized and historical viral **haplotype** network maps for the course of the outbreak are also generated based on all complete and high-quality genomes available. Moreover, 2019nCoVR provides a full collection of SARS-CoV-2 relevant literature on the coronavirus disease 2019 (COVID-19), including published papers from PubMed as well as preprints from services such as bioRxiv and medRxiv through Europe PMC. Furthermore, by linking with relevant **databases** in CNCB, 2019nCoVR offers data submission services for raw sequence reads and assembled genomes, and data sharing with NCBI. Collectively, SARS-CoV-2 is updated daily to collect the latest information on genome sequences, variants, haplotypes, and literature for a timely reflection, making 2019nCoVR a valuable resource for the global research community. 2019nCoVR is accessible at https://bigd.big.ac.cn/ncov/.

Introduction

Coronavirus disease 2019 (COVID-19) is a severe respiratory disease that is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. It has rapidly spread as a pandemic after its outbreak in late December 2019. As of July 14, 2020, 12,964,809 confirmed cases have been reported in 216 countries/territories/areas (WHO Situation Report Number 176; https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/). SARS-CoV-2 samples have been extensively isolated and sequenced by different laboratories across many countries [2], resulting in a considerable number of viral genome sequences worldwide. Therefore, public sharing and free access to a comprehensive collection of SARS-CoV-2 genome sequences is of great significance, which would help to accelerate scientific research and knowledge discovery and also help develop medical countermeasures and sensible decision-making [3].

To date, unfortunately, SARS-CoV-2 genome sequences generated worldwide were scattered around different database resources, primarily including the Global Initiative on Sharing All Influenza Data (GISIAD) [4] repository and NCBI GenBank [5]. Many sequences are available in multiple repositories but their updates are not synchronized. This makes it extremely challenging for worldwide users to effectively retrieve a non-redundant and most updated set of sequence data, and to collaboratively and rapidly deal with this global pandemic. Toward this end, we constructed the 2019 Novel Coronavirus Resource (2019nCoVR, https://bigd.big.ac.cn/ncov/) [6]. Through comprehensive integration and value-added annotation and analysis, we provide public, free, and rapid access to a complete collection of non-redundant global SARS-CoV-2 genomes. Since its inception on January 22, 2020, 2019nCoVR is updated on a daily basis, leading to unprecedentedly dramatic data expansion from 86 genomes in its first release to 64,789 genomes in its current version (as of July 14, 2020). Moreover, it has been substantially upgraded by implementing enhanced data curation and analysis pipelines and online functionalities. Specifically, we enrich 2019nCoVR by including data quality evaluation, variant calling, variant spatiotemporal dynamic tracking, viral haplotype construction, and interactive visualization with more user-friendly web interfaces (**Table 1**). Here we report these significant updates of 2019nCoVR and present the global landscape of SARS-CoV-2 genomes, variants, and haplotypes.

Data collection and processing

Data collection and integration

All genome sequences as well as their related metadata were integrated from SARS-CoV-2 resources worldwide, including

Table 1 Comparison of functional modules between two versions of 2019nCoVR

Functionality	Version 1	Version 2
Integration of coronavirus sequences	✓	✓
Genomic sequences and metadata of 2019-nCoV	✓	✓
Variant statistics and visualization of 2019-nCoV	✓	✓
Phylogenetic tree	✓	✓
Raw sequence data submission	✓	✓
Genome assembly submission	✓	✓
Literature		✓
Clinical information		✓
Sequence integrity and quality assessment		✓
Variant annotation based on protein 3D structures (S protein)		✓
Spatiotemporal dynamics analysis of genomic variants		✓
Haplotype network and dynamic evolution		✓
Artificial intelligence diagnosis and online tools		✓
Enhanced user-friendly interface		✓

GISAID [4], NCBI [5], National Genomics Data Center (NGDC) [7], National Microbiology Data Center (NMDC) [8], and China National GeneBank (CNGB) [9]. To provide a non-redundant dataset, duplicated records from different databases were identified and merged.

Quality control and curation

To ensure the integrity of genome sequences, a sequence is defined as 'complete' if it is longer than 29,000 bp and covers all protein-coding regions of SARS-CoV-2 (nt 266–29674 of GenBank: MN908947.3); otherwise, it is defined as "partial". Furthermore, to examine the quality of genome sequences, unknown bases (Ns) and degenerate bases (Ds, more than one possible base at a particular position and sometimes referred as "mixed bases") were counted for each sequence. According to our definition, a sequence is considered "high-quality" if it contains $\leq$ 15 Ns and $\leq$ 50 Ds, and "low-quality" otherwise. In addition, a sequence is clearly labeled if the number of variants is $\geq$ 15 or the total number of deletions is $\geq$ 2, or the distribution of sequence variations is more aggregated (the ratio of the number of variants divided by the total number of bases in a window is $\geq$ 0.25).

Variant identification and haplotype network construction

Only complete and high-quality genome sequences were used for downstream analyses, including sequence comparison, variant identification, functional annotation, and haplotype network construction. Genome sequence alignment was performed with MUSCLE (3.8.31) [10] by comparing against the earliest released SARS-CoV-2 genome (GenBank: MN908947.3). Sequence variation was identified directly using an in-house Perl program. The effect of variants was determined using Ensembl Variant Effect Predictor (VEP) [11].

SARS-CoV-2 haplotypes were constructed based on short pseudo sequences that consist of all variants (filtering out variations located in UTR regions). Then, all these pseudo sequences were clustered into groups, and each group (a haplotype) represents a unique sequence pattern. The haplotype network was inferred from all identified haplotypes, where the reference sequence haplotype was set as the starting node, and its relationship with other haplotypes was determined according to the inheritance of mutations. As a result, nine major haplotype network clades (denoted as C01–C09) were obtained according to the phylogenetic tree-and-branch structure and the shared landmark mutations (Table 1). Specifically, mutations with population mutation frequency (PMF) $\geq$ 0.05 (except for ATG deletion at position 1605, PMF $\approx$ 0.03) were selected, and the co-occurring mutations were determined by LD linkage analysis. A clade refers to sequences with the co-occurring landmark mutations.

Implementation

2019nCoVR was built based on a browser/server (B/S) architecture. Web interfaces was developed by the Java Server Pages (JSP), HTML, Cascading Style Sheet (CSS), Asynchronous JavaScript and XML (AJAX), JQuery (a cross-platform and feature-rich JavaScript library; http://jquery.com), as well as Semantic-UI (an open source web development framework; https://semantic-ui.com). The database server was implemented by using the Spring Boot (a rapid application development framework based on Spring; https://spring.io). MySQL (https://mysql.com) was used for data storage. For interactive visualization, we implemented HighCharts (a modern SVG-based multi-platform charting library; https://highcharts.com), D3.js (a JavaScript library for manipulating documents based on data; https://d3js.org), and 3Dmol.js (a JavaScript library for visualizing protein structure associated with mutated amino acid residues) [12] in 2019nCoVR. The haplotype network was visualized using D3.js, Leaflet (http://leafletjs.com), and Echarts (http://echarts.baidu.com/).

Database content and features

Statistics of SARS-CoV-2 genome assemblies

Since the outbreak of COVID-19, the number of SARS-CoV-2 genome sequences released globally has been increasing at an unprecedented rate. To facilitate free public access to all genome assemblies and help worldwide researchers better understand the variation and transmission of SARS-CoV-2, we perform daily updates for 2019nCoVR by integrating all available genomes throughout the world and conducting value-added curation and analysis. As of July 14, 2020, 2019nCoVR hosted a total of 64,789 non-redundant genome sequences and provided a global distribution of SARS-CoV-2 genome sequences in 97 countries/regions across 6 continents. Duplicated sequences from different databases are merged with all IDs cross-referenced. Sequences are contributed primarily by United Kingdom (28,823, 44.5%), United States (13,556, 20.9%), Australia (2351, 3.6%), Spain (1852, 2.9%), Netherlands (1605, 2.5%), India (1581, 2.4%), and China (1431, 2.2%). According to our statistics, SARS-CoV-2 genome sequences started to grow rapidly from mid-March (https://bigd.big.ac.cn/ncov/release_genome), concordant with the outset of global pandemic of COVID-19. A full list of our sequence datasets, including strain name, accession number, and source, is provided in Table S1.

To provide high-quality genome sequences that are critically essential for downstream analyses (ranging from variant calling to haplotype construction), we perform sequence integrity and quality assessment for all newly-collected sequences. Among all the human-derived genome sequences released (64,700), 60,970 (94%) are complete, and 31,689 (49%) are high-quality (**Figure 1**A). Most of the low-quality sequences (29,281, 99.7%) contain different numbers of unknown bases (Ns). Among these sequences, 60% have 16–500 Ns (median 258), and 40% have more than 500 Ns (Figure 1B). Further investigation of the genomic locations reveals that some genomic regions with high frequency of Ns (Figure 1C). Sequence integrity and quality assessment analytic data are available for all genome sequences, and can be used as filters for sequence browse and search.

Landscape of genomic variants

Bases on the 31,689 human-derived high-quality complete genome sequences obtained globally (only high-quality complete genome sequences are used for downstream analysis if not indicated otherwise), we investigate the landscape of SARS-

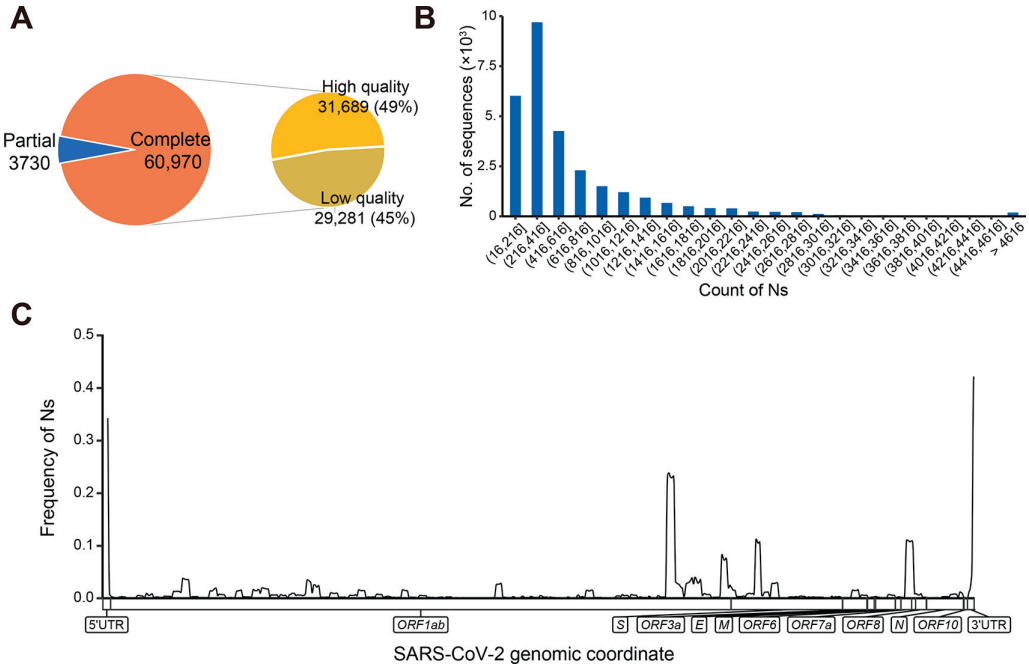

Figure 1 Statistics and distribution of all released SARS-CoV-2 genomes in 2019nCoVR as of July 14, 2020

A. Number and percentage of complete and high-quality genomes. **B.** Distribution of sequence number across different ranges of Ns for low-quality genomes. **C.** Frequency distribution of Ns across the whole genome. A sequence is defined as "complete" if it is longer than 29,000 bp and covers all protein-coding regions of SARS-CoV-2 (nt 266–29674 of GenBank: MN908947.3); otherwise, it is defined as "partial". A sequence is considered "high-quality" if it contains ≤ 15 Ns and ≤ 50 Ds, and "low-quality" otherwise. N, unknown base; D, degenerate base.

CoV-2 genomic variants in comparison with the reference genome (GenBank: MN908947.3) (**Figure 2**). By July 14, 2020, a total of 13,428 variants had been identified, including 12,828 (95.5%) single-nucleotide polymorphisms (SNPs), 437 deletions, 116 insertions, and 47 indels (a combination of an insertion and a deletion, affecting 2 or more nucleotides) (Figure 2A). More than half of these SNPs (6770, 50.4%) are nonsynonymous, causing amino acid changes. To evaluate the impact of missense variants of S protein on the interaction with its receptor human angiotensin-converting enzyme 2 (ACE2) (*e.g.*, in the key binding region), mutated amino acids are projected onto protein 3D structures, which can be viewed by 360 degree rotation (Figure 2B). We further explore distribution of variants across different genes. Noticeably, three genes *ORF1ab*, *S*, and *N* accumulate more variants (Figure 2C). In addition, SNP densities (*i.e.*, number of mutations per nucleotide in the genic region) are higher in several genic regions, including *ORF7a, ORF3a, ORF6*, and *N* (https://bigd.big.ac.cn/ncov/variation/annotation).

For each variant, we investigate its PMF (the ratio of the number of mutated genomes to the total number of complete high-quality genomes) (Figure 2D). Clearly, there are 62 variants with PMF > 0.01 and 18 variants with PMF > 0.05. In particular, there are 4 variants with PMF > 0.75, including positions 241 in 5′UTR, 3037 and 14408 in *ORF1ab*, as well as position 23403 in *S*. These may potentially represent the main prevalent virus genotypes across the globe. All identified variants and their functional annotations are publicly available in the database. In addition, an online pipeline for variant identification and functional annotation is also provided for free access at https://bigd.big.ac.cn/ncov/analysis [13].

Spatiotemporal dynamics of genomic variants

To track the dynamics of SARS-CoV-2 genomic variants, particularly *de novo* mutations, we explore the spatiotemporal change of PMF for each variant according to sampling dates and locations (**Figure 3**). Among the 18 sites with PMF > 0.05, a few mutations occurred simultaneously in multiple sequences and in a linkage manner (Figure 3A), such as mutations at positions 8782 and 28144 as reported previously [14]. It is of note that mutations at these two sites appeared in the early stage of the outbreak on December 30, 2019. Their mutation frequencies reach ~ 0.33 around January 22, 2020, and then gradually declined to 0.10 on July 14, 2020. In contrast, some variants appear only at the middle stage around March 3, 2020. For instance, mutation at position 23403 (resulting in an amino acid change D614G in the S protein) is accompanied by three other mutations, namely, a C-to-U mutation at position 241 in the 5′UTR of SARS-CoV-2 genome, a silent C-to-U mutation in the gene *nsp3* at position 3037, and a missense C-to-U mutation in the gene *RdRp* at position 14408 (P4715L). To make it easier for users to investigate any variant of interest, we provide an interactive heatmap in 2019nCoVR (https://bigd.big.ac.cn/ncov/variation/heatmap) to dynamically display and cluster the mutation patterns over all sampling dates, with customized options available that allow users to select specific variant frequency,

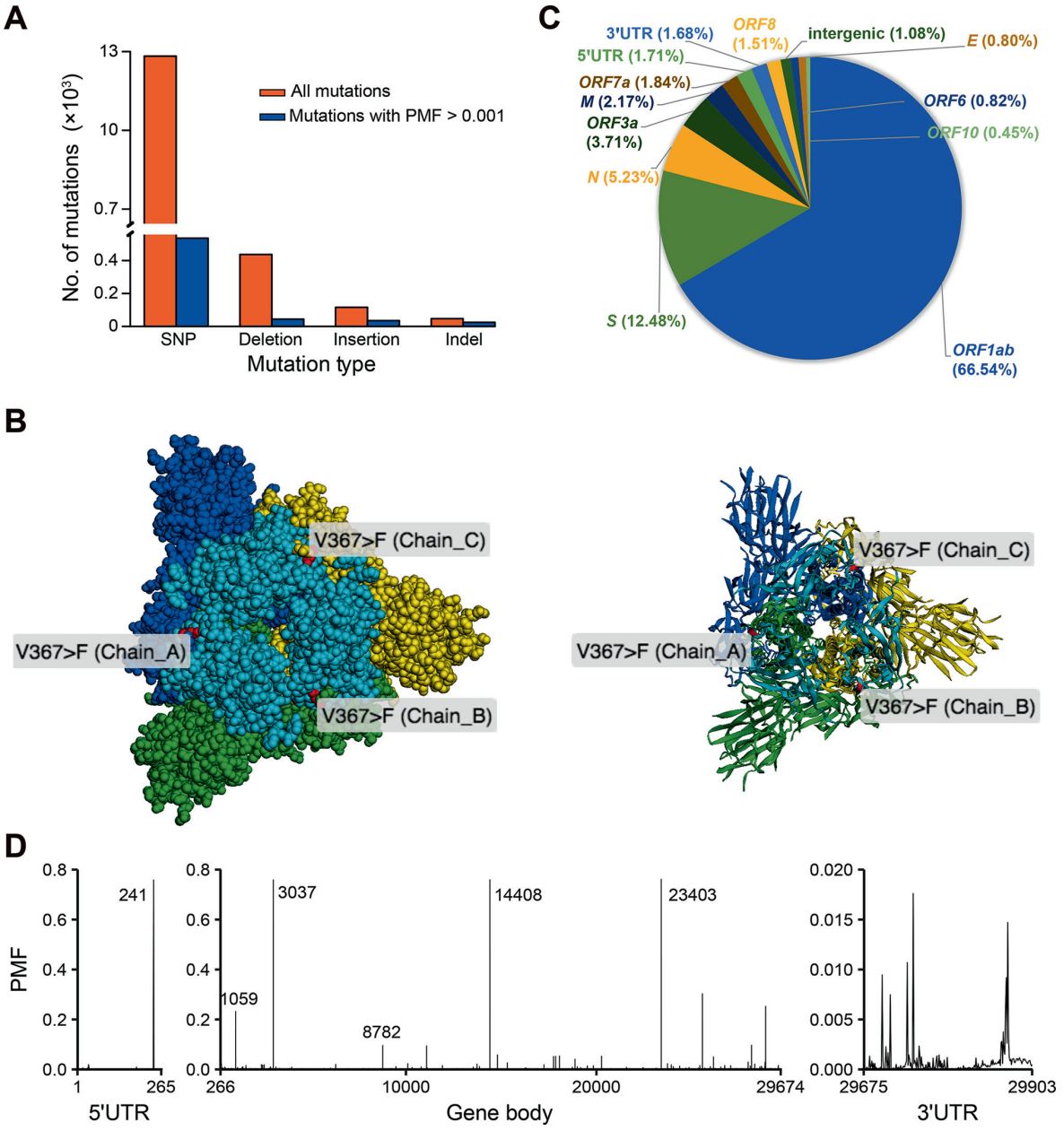

Figure 2 Landscape of genomic variants

A. Number of mutations in different mutation types. The orange bar indicates the number of all mutations, and the blue bar indicates the number of mutations with PMF > 0.001. **B.** 3D structure display for nonsynonymous mutations in S protein (PDB: 6VSB, http://www.rcsb.org/structure/6VSB). The structure is shown in sphere (left panel) and stick (right panel). The three chains (A, B, and C) of S protein are displayed in blue, yellow, and green, respectively. The binding region (amino acid residues: 336–516) of the S protein with its receptor human ACE2 is shown in cyan for all three chains. **C.** Pie chart showing variant annotation for each gene of SARS-CoV-2. **D.** Distribution of PMF for all variants. Coordinate information for representative variants (including positions 241, 1059, 3037, 8782, 14408, and 23403) is provided. PMF, population mutation frequency; ACE2, angiotensin-converting enzyme 2.

annotated gene/region, variant effect type, and transcription regulation sequence (TRS).

Moreover, we investigate dynamic patterns of SARS-CoV-2 genomic variants across different sampling locations over time. Taking the variant at position 23403 (D614G) as an example, its PMF has dramatically increased from 0 at the end of February to 0.76 in the middle of July, and the mutant form G614 became dominant gradually along with the development of pandemic (Figure 3B), presumably indicating that the mutated genotypes may have higher transmissibility [15]. In terms of the absolute number of mutations across different countries/regions, G614 form was dominantly reported in Europe and North America (Figure 3C). (https://bigd.big.ac.cn/ncov/variation/annotation/variant/23403). When investigating the mutation pattern for each country (**Figure 4**), we find that sequences from some Asian countries (such as South Korea,

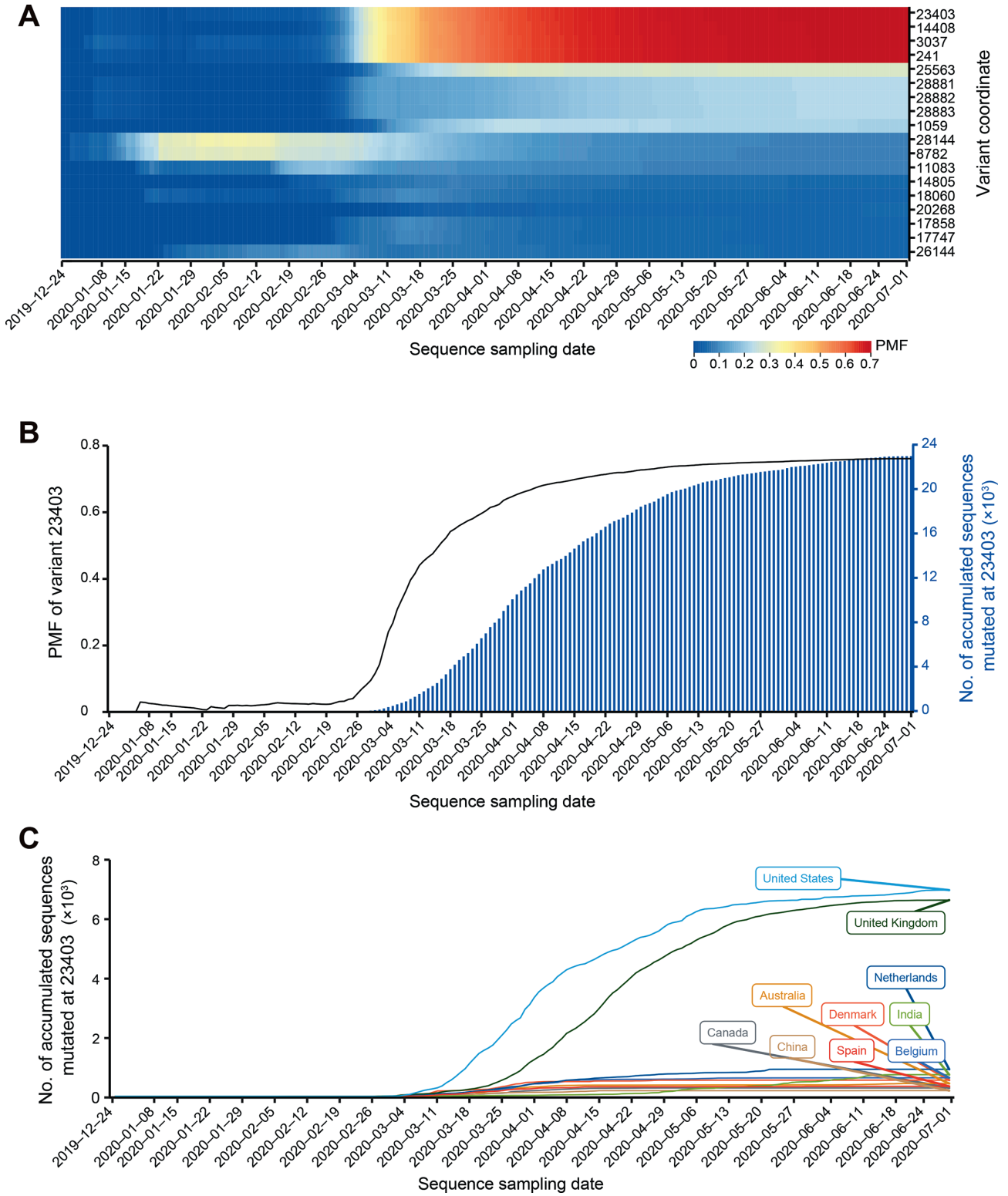

Figure 3 Spatiotemporal dynamics of genomic variants

A. Heatmap of variant PMF (PMF > 0.01) over sampling date. **B.** Distribution of PMF and cumulative growth curve of the sequence with mutation at position 23403 (D614G). **C.** Cumulative growth of the sequence with mutation at position 23403 (D614G) in top 10 countries. Data were downloaded from 2019nCoVR on July 14, 2020.

Malaysia, and Nepal) have no or very few G614 mutant form, whereas countries from Europe and America (*e.g.*, Argentina, Czech Republic, and Serbia) have the G614 form that is dominant among the available samples. In some countries, both the D614 and G614 forms co-existed early in the epidemic, but the mutant form quickly became dominant, such as in Australia,

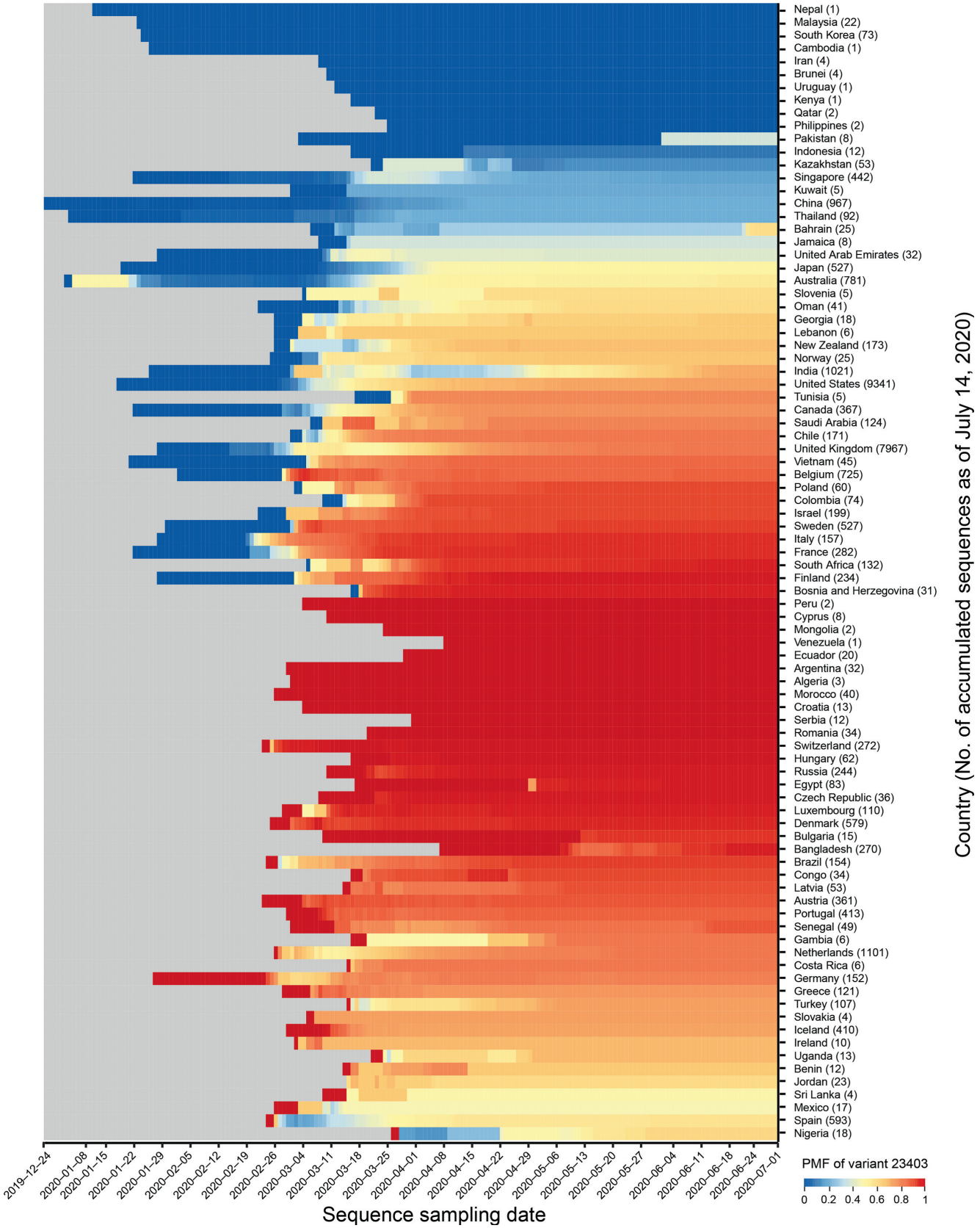

Figure 4 PMF of variant 23,403 for each country across different sampling dates

Number of accumulated sequences as of July 14, 2020 is provided in parenthesis after country name.

Belgium, Canada, Chile, France, Israel, United States, and United Kingdom. The accumulation of this mutation varies in different parts of the world, possibly due to the prevention and control measures implemented in some countries/regions. Taken together, 2019nCoVR features spatiotemporal dynamics tracking of SARS-CoV-2 genomic variants, and thus bears great potential to help decipher viral transmission and adaptation to the host.

Haplotype network construction and characterization

To better characterize the diversity of virus sequences, we built SARS-CoV-2 haplotypes based on all identified variants of non-UTR regions. As a result, 17,624 haplotypes were identified from 31,689 complete high-quality genome sequences as of July 14, 2020. We construct a haplotype network for SARS-CoV-2 (**Figure 5**), a graphical representation of relationships between individual genotypes inferred from genomic variations. The haplotype network is built based on the principle of the shortest set of connections that link all nodes (genotypes), where the length of each connection represents the genetic distance [16]. The SARS-CoV-2 haplotype network can be visualized according to sample collection date and across different countries/regions, thus providing an overview of the pandemic transmission in a spatiotemporal manner. It not only allows users to easily obtain a landscape of SARS-CoV-2 haplotypes and their relationships, but also helps users to navigate a set of haplotypes for a specific country/region. In addition to the haplotype network, the associated information could also be accessed, such as the number of genomes, as well as sampling time and location (Figure 5A).

According to the haplotype network, we classified all genome sequences into nine major clades (labeled as C01–C09; see Methods for details) (Figure 5B and C; **Table 2**). As the pandemic spread of SARS-CoV-2 is still ongoing, new branches that evolve and spread faster are constantly emerging, such as clades C04, C06, C08, and C09 (Table 2). The dominant clades are C06 (8681, 27.4%), C08 (7889, 24.9%), and C09 (6940, 21.9%) (Figure 5D), which are characterized by the signature mutations of C-to-U mutation at positions 241, 3037, and 14408, and A-to-G mutation at position 23403. These clades are defined as the G lineage (as the mutation at position 23403 leads to an amino acid change D614G of S protein). The G lineage sequences have been reported in 82 countries across the globe, and become the main epidemic virus type in most countries in Europe, North America, South America, Africa, and West Asia, *etc*. For example, there are 6827 (71.5%), 8305 (83.4%), and 970 (18.5%) sequences from the G lineage reported in the United States, United Kingdom, and China, respectively (Figure 5E). The widespread and prevalence of the G lineage in different countries suggest the adaptability of this lineage to humans [15].

Discussion

Genome sequencing is vital to understand the epidemiology of SARS-CoV-2, which is not only useful for deciphering genomic composition of the virus and investigating its evolution and transmission, but also highly effective at determining whether individuals belong to the same transmission chain [17]. According to 2019nCoVR, the ratio of the number of sequenced samples to the number of confirmed cases is very low in some countries/regions (Figure S1), and genome sequences are even unavailable in some affected countries/regions. The SARS-CoV-2 sampling bias and limited sequencing depth may lead to inaccurate transmission patterns and phylogenetic relationships [18]. Sequencing all infected cases in a single region reveals that the transmission of *Clostridium difficile* from symptomatic patients accounts for only one third of all infected cases [19]. Given our current understanding of SARS-CoV-2 is still limited, we call for more efforts and collaborations in sequencing more SARS-CoV-2 genomes from both symptomatic and asymptomatic cases.

The SARS-CoV-2 genome sequences currently released were generated by multiple different laboratories on different sequencing platforms. This raises concerns on the quality of genome sequences, such as the Ns of genome, which may affect variant calling and lead to biased population frequency estimation. As mentioned above, the frequency of Ns in some genomic regions is high, possibly due to the low sequencing coverage, low sequence complexity, low efficiency of PCR primers used in sequencing library construction, presence of RNA secondary structure, *etc*. However, sequencing coverage information is largely unavailable, making it challenging to evaluate whether the Ns are due to low sequencing coverage. We further investigated the genomic regions with high frequency of Ns and had the following findings. (1) GC and AG contents of these regions are close to the average GC and AG contents of the whole genome, excluding the possibility of low sequence complexity. (2) The length of these regions ranges from 210 bp to 320 bp (similar to the length of PCR product) and more than 60% of the related sequences are generated on Illumina platform (based on PCR amplification), suggesting that these Ns may result from low efficiency of PCR primers during sequencing library construction. (3) Minimum free energy of RNA secondary structure of these regions is lower than that of randomly extracted regions, indicating that the secondary structure of these regions is more stable and may affect the determination of genome sequences (Figure S2). In future, we plan to construct a golden benchmark dataset with for quality assessment and data filtration.

Compared to the early overly simplified L-S classification [14] and the comprehensive lineages defined by Rambaut et al. [20], our classification scheme with nine clades provides a moderate system that can be correlated with the two classifications mentioned above (Table 2). The nine clades could also be grouped into three lineages defined previously [14], namely, S (C02 and C04), G (C06, C08, and C09), and L (the remaining clades). Although haplotype network cannot provide a precise evolutionary position as phylogenetic trees do, it can be used to quickly inform the clustering of viruses according to signature mutations in each haplotype. Definitely, new clades will be introduced as the virus is continuing to evolve.

A data-driven response to SARS-CoV-2 requires a public, free, and open-access data resource that contains complete high-quality genome sequence data, and equips with automated online pipelines to rapidly analyze genome sequences. Thus, 2019nCoVR (together with other resources in CNCB) provides a wide range of data services, including raw sequencing data archive, genome sequence and meta information management with quality control and curation, variation analysis, as well as data presentation and visualization. Additionally, compared to GISAID and NCBI Virus, 2019nCoVR features

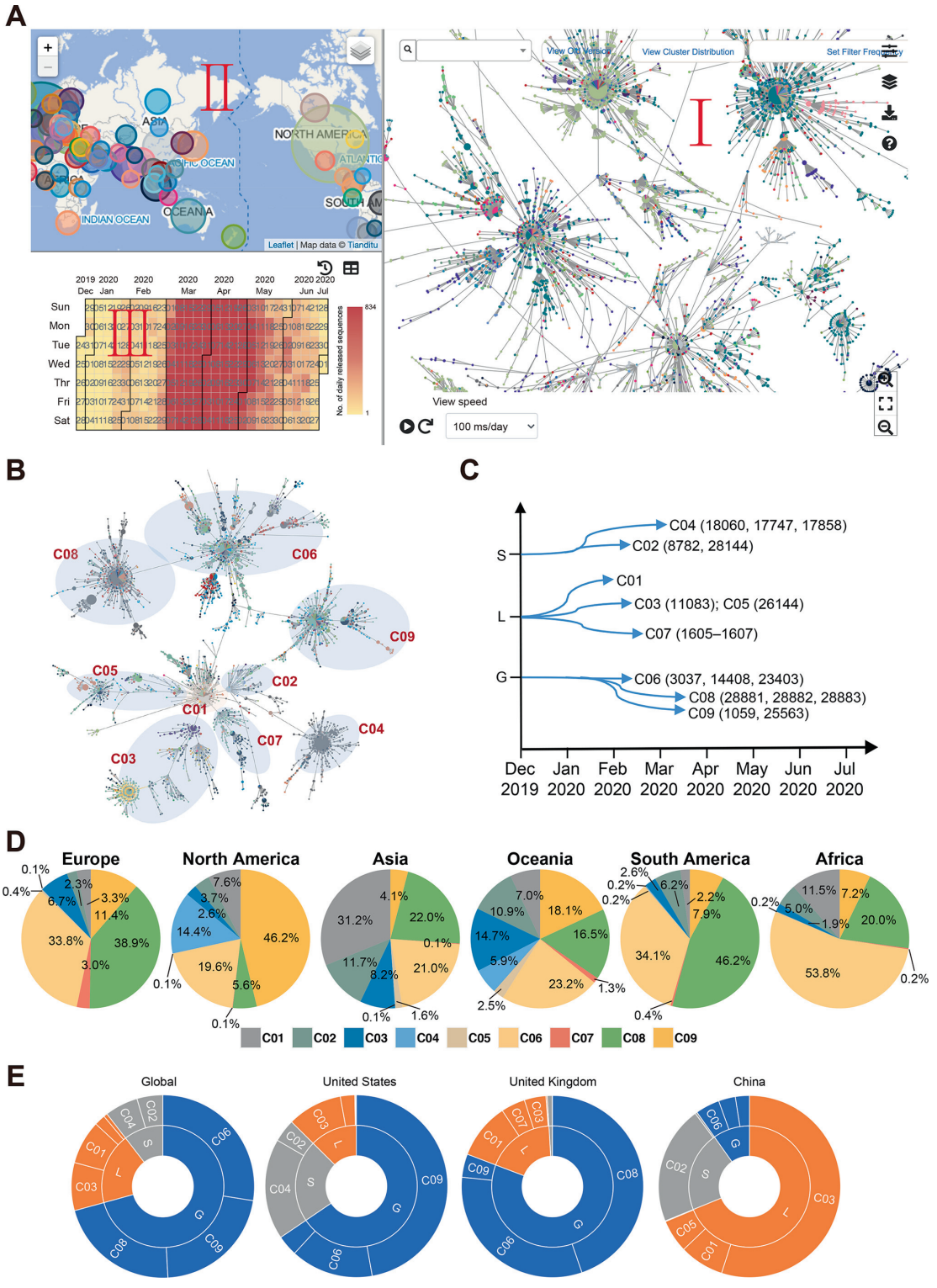

Figure 5 Haplotype network and clade identification and distribution

A. Snapshot of haplotype network dashboard, dynamically showing the development of haplotype (I) across countries (II) and over time (III). Each node in the network represents a haplotype and the node size is proportional to the number of viral genome sequences. The edge between any two nodes represents the genetic distance between two haplotypes. Number of newly-released genome sequences each day is dynamically displayed on the respective date. **B.** Schematic diagram of haplotype clades (C01–C09). **C.** Schematic diagram of three lineages and nine clades, and the common mutation sites for each clade. **D.** Percentage of sequences in clades C01–C09 across different continents. **E.** Sequence number distribution of different lineages (S, L, and G) and clades (C01–C09) throughout the globe and in three representative countries (United States, United Kingdom, and China).

Table 2 Signature mutations of haplotype clades

Lineage	Clade ID	Genomic location	Gene/genomic region	Mutation	AA position and change	Mutation frequency	No. of sequences with the mutation	Classification defined in [7]	Top three lineages defined in [13]
L	C01	NA	NA	NA	NA	NA	NA	L	B/B.3/B.1.3
	C03	11083	*ORF1ab*	G>T	L3606F	0.09	2982	L	B.2/B.2.1/B.4
	C05	26144	*ORF3a*	G>T	G251V	0.05	1592	L	B/B.2
	C07	1604	*ORF1ab*	AATGAC>AAC	ND447N	0.02	503	L	B/B.8
S	C02	8782	*ORF1ab*	C>T	S2839S	0.09	3034	S	A/A.3/A.4/A.5
		28144	*ORF8*	T>C	L84S	0.09	3063		
	C04	17747	*ORF1ab*	C>T	P5828L	0.05	1644	S	A.1/A.1.1/A.1.2
		17858	*ORF1ab*	A>G	Y5865C	0.05	1657		
		18060	*ORF1ab*	C>T	L5932L	0.05	1695		
G	C06	241	*5′UTR*	C>T		0.75	24,028	L	B.1/B.1.5/B.1.11
		3037	*ORF1ab/nsp4*	C>T	F924F	0.75	24,045		
		14408	*ORF1ab/RdRp*	C>T	P4715L	0.75	24,055		
		23403	*S*	A>G	D614G	0.76	24,128		
	C08	28881	*N*	G>A	R203K	0.25	8003	L	B.1/B.1.1/B.1.10
		28883	*N*	G>C	G204R	0.25	7995		
		28882	*N*	G>A	R203R	0.25	7985		
	C09	1059	*ORF1ab*	C>T	T265I	0.23	7357	L	B.1/B.1.21/B.1.43
		25563	*ORF3a*	G>T	Q57H	0.30	9594		

Note: AA, amino acid; NA, not applicable.

spatiotemporal dynamic tracking for all identified variants. This makes it easier for users worldwide to monitor any variant that may be associated with rapid transmission and high virulence. To better understand the epidemiology of SARS-CoV-2, future efforts are needed to collect ever more genome sequences worldwide, to include other types of omics data (such as transcriptome and epitranscriptome, if available) [21], and also to provide more friendly interfaces and online tools in support of research activities worldwide.

Data availability

SARS-CoV-2 genomes, variants (in vcf format), and their annotations are publicly available at https://bigd.big.ac.cn/ncov/.

CRediT author statement

Shuhui Song: Conceptualization, Methodology, Writing - original draft. **Lina Ma:** Data curation, Methodology, Writing - original draft. **Dong Zou:** Resources, Visualization, Writing - original draft. **Dongmei Tian:** Methodology. **Cuiping Li:** Methodology. **Junwei Zhu:** Software. **Meili Chen:** Data curation. **Anke Wang:** Software. **Yingke Ma:** Resources. **Mengwei Li:** Methodology. **Xufei Teng:** Visualization. **Ying Cui:** Data curation. **Guangya Duan:** Data curation. **Mochen Zhang:** Data curation. **Tong Jin:** Data curation. **Chengmin Shi:** Methodology. **Zhenglin Du:** Methodology. **Yadong Zhang:** Methodology. **Chuandong Liu:** Methodology. **Rujiao Li:** Data curation. **Jingyao Zeng:** Data curation. **Lili Hao:** Data curation. **Shuai Jiang:** Methodology. **Hua Chen:** Supervision. **Dali Han:** Supervision. **Jingfa Xiao:** Supervision, Methodology. **Zhang Zhang:** Conceptualization, Supervision, Writing - review & editing. **Wenming Zhao:** Conceptualization, Supervision, Methodology. **Yongbiao Xue:** Conceptualization, Supervision. **Yiming Bao:** Conceptualization, Supervision, Writing - review & editing. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by grants from the Strategic Priority Research Program of Chinese Academy of Sciences (Grant Nos. XDA19090116, XDA19050302, and XDB38030400) awarded to SS, ZZ, and ML; the National Key R&D Program of China (Grant Nos. 2020YFC0848900, 2020YFC0847000, 2016YFE0206600, and 2017YFC0907502); the 13th Five-year Informatization Plan of Chinese Academy of Sciences (Grant No. XXH13505-05); Genomics Data Center Construction of Chinese Academy of Sciences (Grant No. XXH-13514-0202); the Open Biodiversity and Health Big Data Programme of International Union of Biological Sciences, International Partnership Program of Chinese Academy of Sciences (Grant No. 153F11KYSB20160008); the Professional Association of the Alliance of International Science Organizations (Grant No. ANSO-PA-2020-07). This work was also supported by KC Wong Education Foundation to ZZ, as well as the Youth Innovation Promotion Association of Chinese Academy of Sciences (Grant Nos. 2017141 and 2019104) awarded to SS and ML. We thank our colleagues and students for their hard work on the 2019nCoVR (https://bigd.big.ac.cn/ncov). We also thank a number of users and CNCB members for report-

ing bugs and sending comments. Complete genome sequences used for analyses were obtained from the Genome Warehouse of CNCB, CNGBdb, GenBank, GISAID, and NMDC resources. We acknowledge the sample providers and data submitters listed in Table S1.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2020.09.001.

ORCID

0000-0003-2409-8770 (Shuhui Song)
0000-0001-6390-6289 (Lina Ma)
0000-0002-7169-4965 (Dong Zou)
0000-0003-0564-625X (Dongmei Tian)
0000-0002-7144-7745 (Cuiping Li)
0000-0003-4689-3513 (Junwei Zhu)
0000-0003-0102-0292 (Meili Chen)
0000-0002-2565-2334 (Anke Wang)
0000-0002-9460-4117 (Yingke Ma)
0000-0001-6163-2827 (Mengwei Li)
0000-0001-9282-4282 (Xufei Teng)
0000-0001-9201-0465 (Ying Cui)
0000-0003-4582-5156 (Guangya Duan)
0000-0001-9136-451X (Mochen Zhang)
0000-0003-0791-2822 (Tong Jin)
0000-0003-0237-4092 (Chengmin Shi)
0000-0003-2147-3475 (Zhenglin Du)
0000-0003-0918-5673 (Yadong Zhang)
0000-0002-9904-7786 (Chuandong Liu)
0000-0002-3276-8335 (Rujiao Li)
0000-0001-7364-9677 (Jingyao Zeng)
0000-0003-3432-7151 (Lili Hao)
0000-0002-6722-176X (Shuai Jiang)
0000-0002-9829-6561 (Hua Chen)
0000-0001-7119-1578 (Dali Han)
0000-0002-2835-4340 (Jingfa Xiao)
0000-0001-6603-5060 (Zhang Zhang)
0000-0002-4396-8287 (Wenming Zhao)
0000-0002-6895-8472 (Yongbiao Xue)
0000-0002-9922-9723 (Yiming Bao)

References

[1] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 2020;5:536–44.
[2] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579:265–9.
[3] Zhang Z, Song S, Yu J, Zhao W, Xiao J, Bao Y. The elements of data sharing. Genomics Proteomics Bioinformatics 2020;18:1–4.
[4] Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data - from vision to reality. Euro Surveill 2017;22:30494.
[5] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733–45.
[6] Zhao WM, Song SH, Chen ML, Zou D, Ma LN, Ma YK, et al. The 2019 novel coronavirus resource. Yi Chuan 2020;42:212–21. (in Chinese with an English abstract)
[7] National Genomics Data Center Members and Partners. Database resources of the National Genomics Data Center in 2020. Nucleic Acids Res 2020;48:D24–33.
[8] Shi W, Qi H, Sun Q, Fan G, Liu S, Wang J, et al. gcMeta: a global catalogue of metagenomics platform to support the archiving, standardization and analysis of microbiome data. Nucleic Acids Res 2019;47:D637–48.
[9] Xiao SZ, Armit C, Edmunds S, Goodman L, Li P, Tuli MA, et al. Increased interactivity and improvements to the GigaScience database GigaDB. Database (Oxford) 2019;2019:1–9.
[10] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7.
[11] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. Genome Biol 2016;17:122.
[12] Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. Bioinformatics 2015;31:1322–4.
[13] Gong Z, Zhu JW, Li CP, Jiang S, Ma LN, Tang BX, et al. An online coronavirus analysis platform from the National Genomics Data Center. Zool Res 2020;41:705–8.
[14] Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 2020;7:1012–23.
[15] Korber B, Fischer W, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020;182:812–27.e19.
[16] Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 1999;16:37–48.
[17] Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. Curr Opin Microbiol 2015;23:62–7.
[18] Mavian C, Pond SK, Marini S, Magalis BR, Vandamme AM, Dellicour S, et al. Bias and incorrect rooting make phylogenetic network tracing of SARS-CoV-2 infections unreliable. Proc Natl Acad Sci U S A 2020;117:12522–3.
[19] Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, et al. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. N Engl J Med 2013;369:1195–205.
[20] Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5:1403–7.
[21] Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. Cell 2020;181:914–21.e10.