The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR

Shuhui Song[1,2,3,4,#,a], Lina Ma[1,2,3,#,b], Dong Zou[1,2,3,#,c], Dongmei Tian[1,2,#,d], Cuiping Li[1,2,#,e], Junwei Zhu[1,2,#,f], Meili Chen[1,2,3,g], Anke Wang[1,2,h], Yingke Ma[1,2,i], Mengwei Li[1,2,3,4,j], Xufei Teng[1,2,3,4,k], Ying Cui[1,2,3,4,l], Guangya Duan[1,2,3,4,m], Mochen Zhang[1,2,3,4,n], Tong Jin[1,2,3,4,o], Chengmin Shi[1,5,p], Zhenglin Du[1,2,3,q], Yadong Zhang[1,2,3,4,r], Chuandong Liu[1,5,s], Rujiao Li[1,2,3,t], Jingyao Zeng[1,2,3,u], Lili Hao[1,2,3,v], Shuai Jiang[1,2,w], Hua Chen[1,5,x], Dali Han[1,5,y], Jingfa Xiao[1,2,3,4,z], Zhang Zhang[1,2,3,4,*,aa], Wenming Zhao[1,2,3,4,*,ab], Yongbiao Xue[1,2,4,*,ac], Yiming Bao[1,2,3,4,*,ad]

[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] Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

[#]These authors contributed equally to this work.

[*]Corresponding author (baoym@big.ac.cn, ybxue@big.ac.cn, zhaowm@big.ac.cn zhangzhang@big.ac.cn,)

Running title: *Song S et al /landscape of SARS-CoV-2 genomes, variants, and haplotypes*

[a]ORCID: 0000-0003-2409-8770

[b]ORCID: 0000-0001-6390-6289

[c]ORCID: 0000-0002-7169-4965

[d]ORCID: 0000-0003-0564-625X

[e]ORCID: 0000-0002-7144-7745

[f]ORCID: 0000-0003-4689-3513

[g]ORCID: 0000-0003-0102-0292

[h]ORCID: 0000-0002-2565-2334

[i]ORCID: 0000-0002-9460-4117

[j]ORCID: 0000-0001-6163-2827

[k]ORCID: 0000-0001-9282-4282

[l]ORCID: 0000-0001-9201-0465

[m]ORCID: 0000-0003-4582-5156

[n]ORCID: 0000-0001-9136-451X

[o]ORCID: 0000-0003-0791-2822

[p]ORCID: 0000-0003-0237-4092

[q]ORCID: 0000-0003-2147-3475

[r]ORCID: 0000-0003-0918-5673

[s]ORCID: 0000-0002-9904-7786

[t]ORCID: 0000-0002-3276-8335

[u]ORCID: 0000-0001-7364-9677

[v]ORCID: 0000-0003-3432-7151

[w]ORCID: 0000-0002-6722-176X

[x]ORCID: 0000-0002-9829-6561

[y]ORCID: 0000-0001-7119-1578

[z]ORCID: 0000-0002-2835-4340

[aa]ORCID: 0000-0001-6603-5060

[ab]ORCID: 0000-0002-4396-8287

[ac]ORCID: 0000-0002-6895-8472

[ad]ORCID: 0000-0002-9922-9723

Total word counts: 4477

Total figures: 5

Total tables: 1

Total supplementary figures and tables: 2 (1 Figure S1, 1 Table S1)

Abstract

On 22 January 2020, the China National Center for Bioinformation (CNCB) / National Genomics Data Center (NGDC) created the 2019 Novel Coronavirus Resource (2019nCoVR, https://bigd.big.ac.cn/ncov/), an open-accessed SARS-CoV-2 information resource. 2019nCoVR features 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 our in-house automated pipeline. Of particular note, 2019nCoVR performs systematic analyses to obtain a dynamic landscape of SARS-CoV-2 genomic variations at a global scale. It provides all identified variants and detailed statistics for each virus isolate, and congregates the quality score, functional annotation, and population frequency for each variant. It also generates visualization of the spatiotemporal change for each variant and yields historical viral haplotype network maps for the course of the outbreak from all complete and high-quality genomes. Moreover, 2019nCoVR provides a full collection of literatures on 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/NGDC, 2019nCoVR offers data submission services for raw sequence reads and assembled genomes, and data sharing with National Center for Biotechnology Information. Collectively, all SARS-CoV-2 genome sequences, variants, haplotypes and literatures are updated daily to provide timely information, making 2019nCoVR a valuable resource for the global research community.

KEYWORDS: 2019nCoVR; SARS-CoV-2; Genome assembly; Genomic variation; Haplotype

Introduction

The severe respiratory disease COVID-19 [1], since its outbreak in late December 2019, has rapidly spread as a pandemic. As of 14 July 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/). As the causative agent of COVID-19, SARS-CoV-2 samples have been extensively isolated and sequenced by different countries and laboratories [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 for worldwide researchers 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 exist 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. Towards this end, we constructed the 2019 novel coronavirus resource (2019nCoVR, https://bigd.big.ac.cn/ncov/) in the China National Center for Bioinformation (CNCB) / National Genomics Data Center (NGDC), with the purpose to provide public, free, rapid access to a complete collection of non-redundant global SARS-CoV-2 genomes by comprehensive integration and value-added annotation and analysis [6]. Since its inception on 22 January 2020, 2019nCoVR is updated on 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 14 July 2020). Moreover, it has been greatly upgraded by equipping with enhanced data curation and analysis pipelines and online functionalities, including data quality evaluation, variant calling, variant spatiotemporal dynamic tracking, viral haplotype construction, and interactive visualization with more friendly web interfaces. Here we report these significant updates of 2019nCoVR.

Results

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 public free 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 (Figure 1). As of 14th July 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 (Figure 1A). 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 (2,351, 3.6%), Spain (1,852, 2.9%), Netherlands (1,605, 2.5%), India (1,581, 2.4%), and China (1,431, 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 dataset 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 released human-derived genome sequences (64,700), 60,970 (94.2%) are complete, and 31,689 (49%) are high-quality (with high coverage) (Figure 1B). 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 1C). Further investigation of the genomic locations reveals that some genomic regions have high coverage of Ns (Figure 1D). Sequence integrity and quality assessment results are available for all genome sequences, and can be used as filters for sequence browse and search.

Landscape of genomic variants

Bases on 31,685 globally human-derived high-quality complete genome sequences (in what follows, only high-quality complete genome sequences are used for downstream analysis if not indicated otherwise), we investigate the landscape of SARS-CoV-2 genomic variants by comparison with the reference genome (MN908947.3 in NCBI) (Figure 2). By 14th July 2020, a total of 13,428 variants were 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 (6,770, 50.4%) are nonsynonymous, causing amino acid changes. To gain the functional effects of those missense variants from the perspective of spatial location (e.g. key functional domain or binding region), mutated amino acids are projected onto protein 3D structures, which can be viewed by 360 degree rotation (Figure 2B). We further explore the distributions of these variants across different genes. Noticeably, the three genes *ORF1ab*, *S*, and *N* accumulate more variants (Figure 2C) and SNP densities (i.e., the number of mutations per nucleotides in the gene region) are higher in several gene regions including *ORF7a*, *ORF3a*, *ORF6 and N*.

For each variant, we investigate its population mutation frequency (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 > 1%, 18 variants with PMF > 5%, and 4 variants with PMF > 75.8% (that is, position 241 in 5'UTR, positions 3,037 and14,408 in *ORF1ab*, and 23,403 in *S*), potentially representing main prevalent virus genotypes across the global. All identified variants and their functional annotations are publicly accessible at https://bigd.big.ac.cn/ncov/variation/annotation, and an online pipeline for variant identification and functional annotation is provided and freely available at https://bigd.big.ac.cn/ncov/analysis.

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 population frequency for each variant according to sampling time and location (Figure 3). Among the 18 sites with PMF > 5%, some are mutated simultaneously and in a linkage manner (Figure 3A), such as mutations at positions 8,782 and 28,144 reported in [7]. Specifically, these two sites appeared in the early stage of the outbreak since 30 December 2019, and their mutation frequencies reach ~33% around 22 January 2020, and then gradually decline to 9.6% currently. Contrastingly, several variants appear only since the middle stage around 3 March 2020; such as the mutation at position 23,403 (provoking an amino acid change D614G of the S protein), is accompanied by three other mutations, namely, a C-to-U mutation at position 241 in the 5'UTR, a silent C-to-U mutation in the gene *nsp3* at position 3,037, and a missense C-to-U mutation in the gene *RdRp* at position 14,408 (P4715L). To facilitate 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, 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 23,403 (D614G) as an example, its PMF has dramatically increased from 0 at the end of February to 76.2% right now, and the mutation pattern G614 has been gradually dominant along with the development of pandemic (Figure 3B), presumably indicating that the mutated genotypes may have higher transmissibility[8]. In terms of the absolute number of mutation patterns across different countries/regions, G614 emerges dominantly 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, Malaysia, and Nepal) have no or very few G614 mutation, whereas those from Europe and America (e.g. Argentina, Czech Republic and Serbia) do have the G614 pattern that is dominated among contemporary samples. In some countries, both the D614 and G614 patterns are co-circulating early in the epidemic, but the mutated pattern soon begins to be dominant such as in Australia, Belgium, Canada, Chile, France, Israel, United States and United Kingdom [8]. The accumulation of this mutation varies in different parts of the world, possibly due to the prevention and control measures adopted by 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 [8].

Haplotype network construction and characterization

To better characterize the diversity of virus sequences, we build SARS-CoV-2 haplotypes based on all identified variants beyond UTRs regions. As a result, 17,624 haplotypes were identified from 31,685 complete high-quality genome sequences as of 14th July 2020. Based on this, we construct a haplotype network for SARS-CoV-2 (Figure 5), a graphical representation of genomic variations by inferring relationships between individual genotypes, according to the principle of the shortest set of connections that link all nodes (genotypes) where the length of each connection represents the genetic distance [9]. To provide a whole picture of the pandemic transmission in a spatiotemporal manner, we visualize the SARS-CoV-2 haplotype network by sample collection date and across different countries/regions. 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 linking with additional associated information such as the number of genomes, sampling time and location (Figure 5A).

According to the haplotype network, we classify all genome sequences into nine major clusters (labelled as C01-C09; see Methods for detail) (Figure 5B, 5C; Table 1). As the ongoing pandemic spread of SARS-CoV-2, new branches that evolve and spread faster are gradually emerging, such as clusters C04, C06, C08 and C09 (Table 1). The dominant clusters are C06 (8,681, 27.4%), C08 (7,889, 24.9%) and C09 (6,940, 21.9%) (Figure 5D), which are characterized by those signature mutations of C241T, C3037T, C14408T, and A23403G, and are defined as the G clade (as the mutation at position 23,403 provoking an amino acid change D614G of S protein). These sequences have spread to 82 countries worldwide, 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 about 6,827 (71.5%), 8,305 (83.4%) and 970 (18.5%) sequences originated from the G clade in United States, United Kingdom, and China, respectively (Figure 5E). The wide spread and prevalence of this clade in different countries suggests the adaptability of the virus type to human [8].

Discussion

Whole genome sequencing is vital to deal with SARS-CoV-2, since it is not only useful for deciphering its genome sequences and investigating its evolution and transmission, but also highly effective at determining whether individuals are part of the same transmission chain [10]. It has proved that sequencing all cases of infection reported in a single region show that there are other routes of transmission than solely between symptomatic patients [11]. According to 2019nCoVR, however, the density of sampling, compared with the number of confirmed cases, is very low in some countries/regions (Figure S1), and even genome sequences are unavailable in some affected countries/regions. The SARS-CoV-2 sampling bias and depth may lead to inaccurate transmission patterns and phylogenetic relationships [12]. As our current understanding is still very limited, we call for more efforts and collaborations in sequencing more SARS-CoV-2 genomes from both patients and asymptomatic infections. Besides, as those released SARS-CoV-2 genome sequences are generated by multiple different laboratories on different sequencing platforms, the quality of genome sequences is another important factor, which may affect variant calling and population frequency estimation. As mentioned, sequencing coverage of SARS-CoV-2 is diverse, which may lead to biased genotype frequency for some mutation sites. Our future efforts are to construct a recognized benchmark for quality assessment and data filtration.

Compared to the early overly simplified L-S classification [8] and those comprehensive lineages defined by Rambaut et al. [10][13], our classification scheme with nine clusters provides a moderate system that can be correlated with the others (Table 1). The nine clusters could also be grouped into three clades defined in [8, 10], S (C02 and C04), G (C06, C08 and C09), and L (the rest clusters). Although haplotype network cannot give a precise evolutionary position as phylogenetic trees do, it can be used to quickly obtain the clustering of viruses according to signature mutations in each haplotype. New clusters will be introduced as the virus continue 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 is equipped with automated pipelines to rapidly analyze genome sequences in real-time. Thus, 2019nCoVR (as well as other resources in NGDC/CNCB) provides a wide range of data services, involving raw sequencing data archive, genome sequence and meta information management with quality control and curation, variation analysis and data presentation and visualization. Additionally, 2019nCoVR provides spatiotemporal dynamic tracking for all identified variants in order to facilitate worldwide users to monitor any variant that may be associated with rapid transmission and high virulence. To better understand the life cycle and pathogenicity of SARS-CoV-2, future directions are to collect more and more genome sequences worldwide, include other types of omics data (such as transcriptome and epitranscriptome, if available) [14] and also provide more friendly interfaces and online tools in support of worldwide research activities.

Methods

Data collection and integration

All genome sequences as well as their related metadata were integrated from SARS-CoV-2 resources worldwide, including NCBI [5], GISAID [15], CNCB/NGDC [16], NMDC [17] and CNGB [18]. To provide a non-redundant dataset, duplicated records across different databases were identified and merged.

Quality control and curation

To determine the integrity of genome sequences, one sequence is defined as ‘Complete’ if it has ≧29000 bases and covers all protein-coding/CDS regions of SARS-CoV-2 (bases 266:29674 of NCBI entry 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. By our default definition, one sequence is “high-quality” if Ns≦15 and Ds≦50, and “low-quality” otherwise. Besides, any sequence is clearly labelled when the number of variants≧15 or the total number of deletion≧2 or the distribution of sequence variation is more aggregated (the ratio of the number of variants divided by the total number of bases in a window≧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) [19] by comparing against the earliest released SARS-CoV-2 genome (MN908947.3). Sequence variation was identified directly using an in-house Perl program. The effect of variants was determined using VEP (ENSEMBL Variant Effect Predictor) [20].

SARS-CoV-2 haplotypes were constructed based on short pseudo sequences that consist of all variants (filtering out variations located in UTR regions) only. Then, all these pseudo sequences were clustered into groups, and each group (a haplotype) represents a unique sequence pattern. The haplotype network is inferred from all identified haplotypes, where the reference sequence haplotype is set as the starting node, and its relationship with other haplotypes is determined according to the inheritance of mutations. As a result, nine major haplotype network clusters (denoted as C01-C09) are obtained according to the phylogenetic tree-and-branch structure and those shared landmark mutations (Table 1). Specifically, mutations with PMF≧5% (except for ATG deletion at position1605, PMF≈3%) were selected, and those co-occurred mutations were determined by LD linkage analysis. A cluster is referred to sequences with those co-occurred landmark mutations.

Database construction and visualization

2019nCoVR was built based on B/S (Browser/Server) architecture. In the browser-side, it was developed by JSP (Java Server Pages), HTML, CSS, AJAX (Asynchronous JavaScript and XML), 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). In the server-side, it was implemented by using Spring Boot (a rapid application development framework based on Spring; https://spring.io). For data storage, MySQL (https://mysql.com) was used. For interactive visualization, 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 acids) [21] were employed in 2019nCoVR. The haplotype network was implemented by d3js, Leaflet (http://leafletjs.com), and Echarts (http://echarts.baidu.com/).

Data availability

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

Authors' contributions

YB, YX, WZ and ZZ designed, supervised, and coordinated the study. SS and JX participated in the design of the study. LM, YC, GD, MZ, TJ and MC performed data curation and download. SS, DT and CL analyzed data. SS, XT and CL drew the figures. DZ, JZ, AW, YM, ML and YZ developed database. SS, DT, CL, ZD, SJ and JZ developed the online analysis tools. RL, JZ and LH involved in literature reports. HC, DH, and JX involved in haplotype and 3D structure. SS and ZZ drafted the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Funding

This work was supported by grants from The Strategic Priority Research Program of the Chinese Academy of Sciences [XDA19090116 to S.S., XDA19050302 to Z.Z., XDB38030400 to L.M.], National Key R&D Program of China [2020YFC0848900, 2016YFE0206600, 2017YFC0907502], 13th Five-year Informatization Plan of Chinese Academy of Sciences [XXH13505-05], Genomics Data Center Construction of Chinese Academy of Sciences [XXH-13514-0202], K. C. Wong Education Foundation to Z.Z., and International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008]. The Youth Innovation Promotion Association of Chinese Academy of Science [2017141 to S.S., 2019104 to L.M.]; Funding for open access charge: The

Strategic Priority Research Program of the Chinese Academy of Sciences.

Acknowledgements

We thank our colleagues and students for their hard working on the 2019nCoVR (https://bigd.big.ac.cn/ncov). We also thank a number of users and CNCB/NGDC members for reporting bugs and sending comments and particularly all sample collection and data submission units/individuals.

References

[1] Coronaviridae Study Group of the International Committee on Taxonomy of V. 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.

[4] Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill 2017;22.

[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.

[7] Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. National Science Review 2020.

[8] Korber B, Fischer W, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-22020:2020.04.29.069054.

[9] Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 1999;16:37-48.

[10] Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. Curr Opin Microbiol 2015;23:62-7.

[11] 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.

[12] Mavian C, Pond SK, Marini S, Magalis BR, Vandamme AM, Dellicour S, et al. Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. Proc Natl Acad Sci U S A 2020.

[13] 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.

[14] 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.

[15] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 2017;1:33-46.

[16] National Genomics Data Center M, Partners. Database Resources of the National Genomics Data Center in 2020. Nucleic Acids Res 2020;48:D24-D33.

[17] 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-D48.

[18] 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.

[19] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792-7.

[20] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122.

[21] Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. Bioinformatics 2015;31:1322-4.

Figure legends

Figure 1 Statistics and distribution of all open released SARS-CoV-2 genomes. (a) Distribution of released genome sequences by country, territory or region; (b) Number and percentage of complete and high-quality genomes; (c) Sequence numbers of different Ns ranges for low-quality genomes; (d) Distribution of Ns across the whole genome.

Figure 2 Landscape of genomic variants. (a) Numbers of different mutation types (including SNPs, deletion, insertion and indels); (b) Pie chart of variant annotation for each gene; (c) Structure display for nonsynonymous mutations; (d) Mutation frequencies for all variants.

Figure 3 Spatiotemporal dynamics of genomic variants. (a) Heatmap shows the population mutation frequency (PMF) of variants over time; an example of dynamic PMF and cumulative sequence growth curve (b) and the cumulative growth curve of the number of mutated viruses in each country (c) for position (n23403, pD614G).

Figure 4 The population mutated frequency (PMF) of G614 for each country over time.

Figure 5 Haplotype network and cluster identification and distribution. (a) The snapshot of haplotype network dashboard, which can dynamically show the development of haplotype over time and country. Each node in the network represents a haplotype and the node size is proportional to the number of viral genome sequences, where the edge between any two

nodes represents the genetic distance between two haplotypes (i.e. the number of mutation sites); (b-c) Schematic diagram of haplotype clusters (C01-C09) and their corresponding common mutation sites for each cluster; (d) Distribution of C01-C09 clusters across different continents; (e) Distribution of different clusters throughout the world, and in three representative countries (US, UK and China).

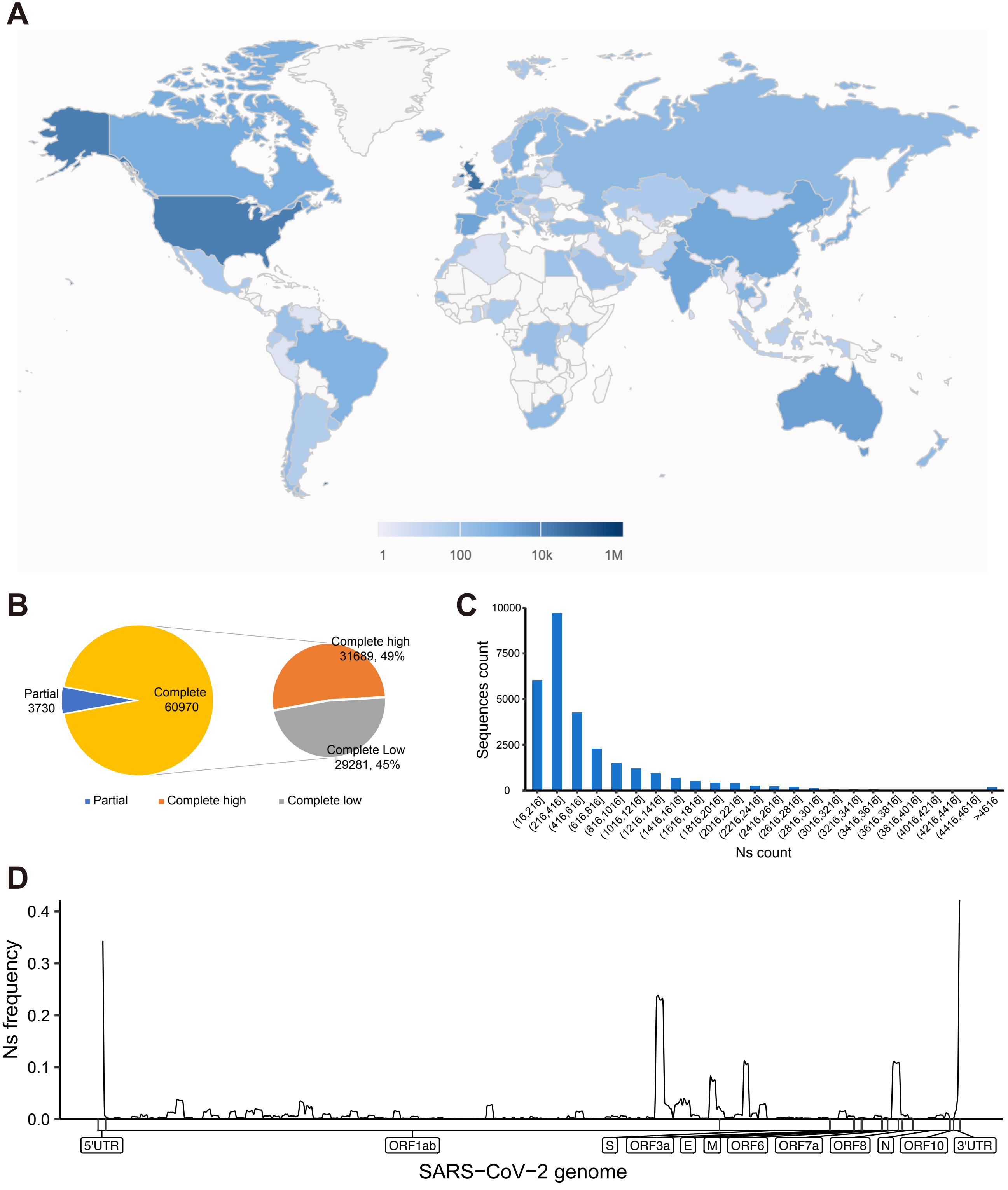
Supplementary information

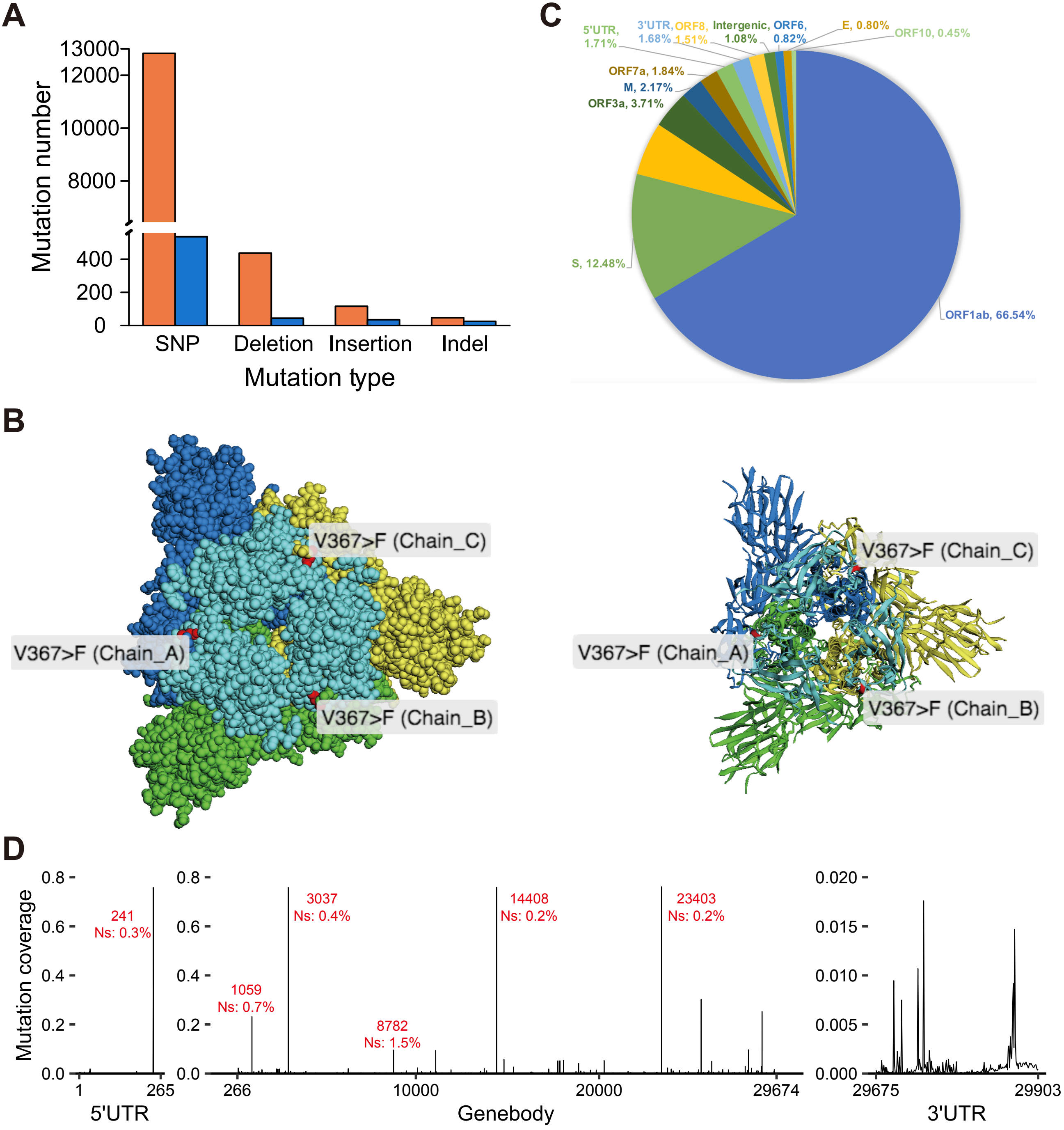
Figure S1 The distribution of genome sequence count divided by the number of confirmed cases for each country.

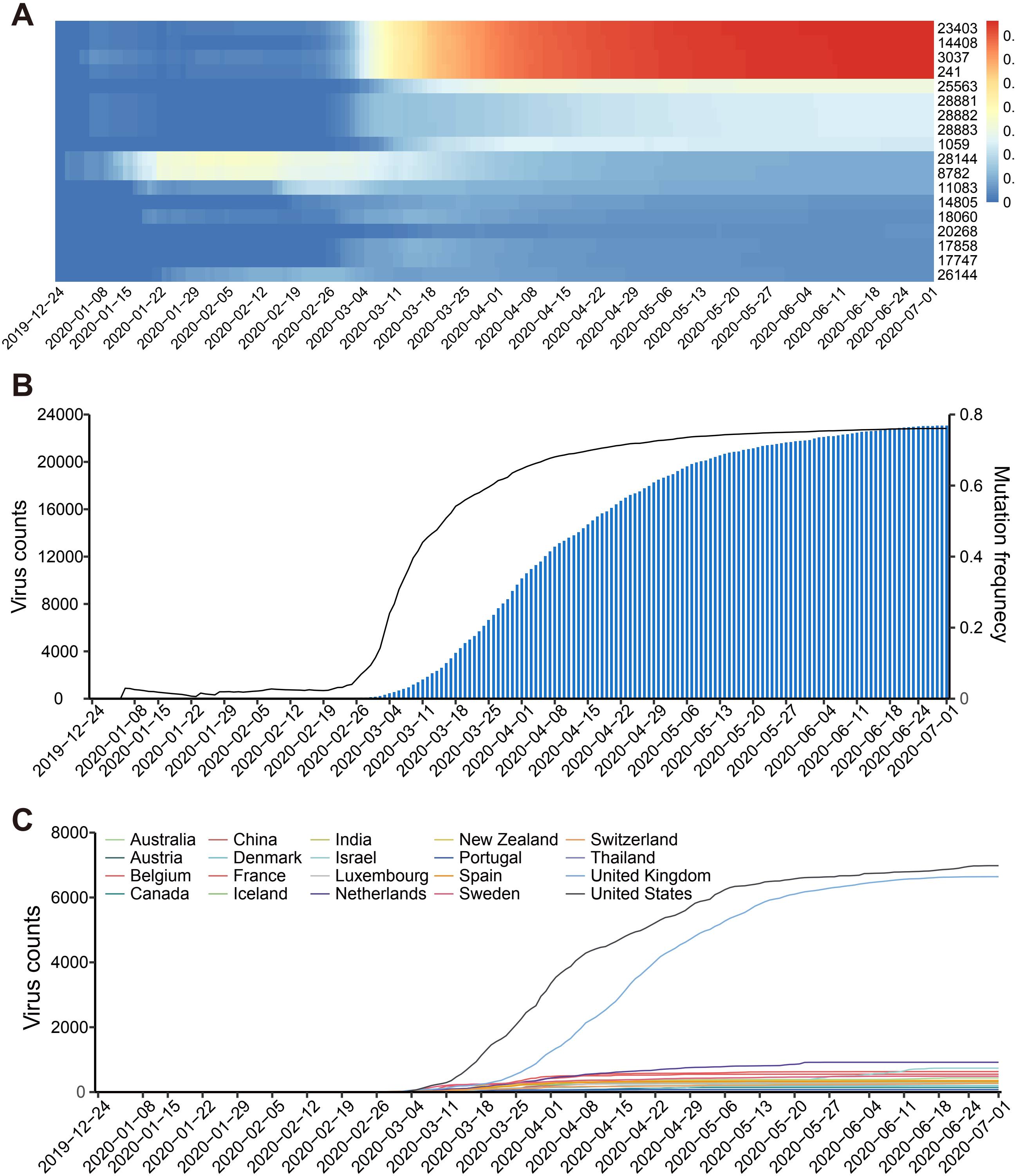
Table. S1 Coronavirus sequence datasets used for the study.

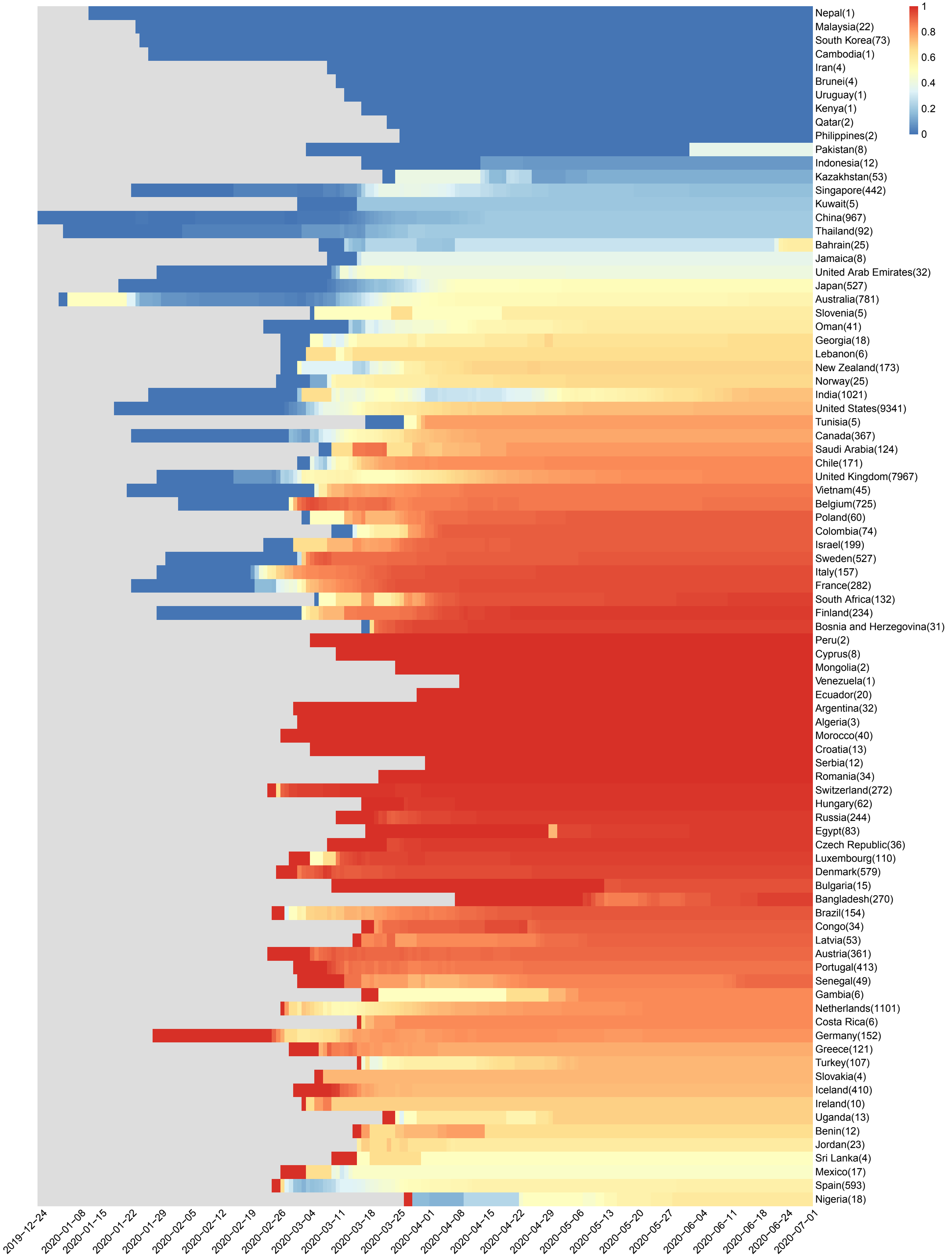
Table 1 The signature mutations for haplotype clusters

Clade	Cluster ID	Genomic Position	Gene/Regions	Nuc Change	Protein Position and Change	Mutation Frequency	Mutated Numbers	Lu et.al 2020 NSR	Rambaut et.al 2020 Nat Microbiol
L	C01	NA	NA	NA	NA	NA	NA	L	B/B.3/B.1.3 etc.
	C03	11083	orf1ab	G->T	p.3606L>F	0.09	2982	L	B.2/B.2.1/B.4 etc.
	C05	26144	ORF3a	G->T	p.251G>V	0.05	1592	L	B/B.2
	C07	1604	orf1ab	AATG->A	p.447-448ND>N	0.02	503	L	B/B.8
S	C02	8782	orf1ab	C->T	p.2839S	0.09	3034	S	A/A.3/A.4/A.5
		28144	ORF8	T->C	p.84L>S	0.09	3063		
	C04	17747	orf1ab	C->T	p.5828P>L	0.05	1644	S	A.1/A.1.1/A.1.2
		17858	orf1ab	A->G	p.5865Y>C	0.05	1657		
		18060	orf1ab	C->T	p.5932L	0.05	1695		
G	C06	241	5'UTR	C->T		0.75	24028	L	B.1/B.1.5/B.1.11 etc.
		3037	orf1ab/nsp4	C->T	p.924F	0.75	24045		
		14408	orf1ab/RdRp	C->T	p.4715P>L	0.75	24055		
		23403	S	A-> G	p.614D>G	0.76	24128		
	C08	28881	N	G->A	p.203R>K	0.25	8003	L	B.1/B.1.1/B.1.10 etc.
		28883	N	G-> C	p.204G>R	0.25	7995		
		28882	N	G->A	p.203R	0.25	7985		
	C09	1059	orf1ab	C->T	p.265T>I	0.23	7357	L	B.1/B.1.21/B.1.43
		25563	ORF3a	G->T	p.57Q>H	0.30	9594		

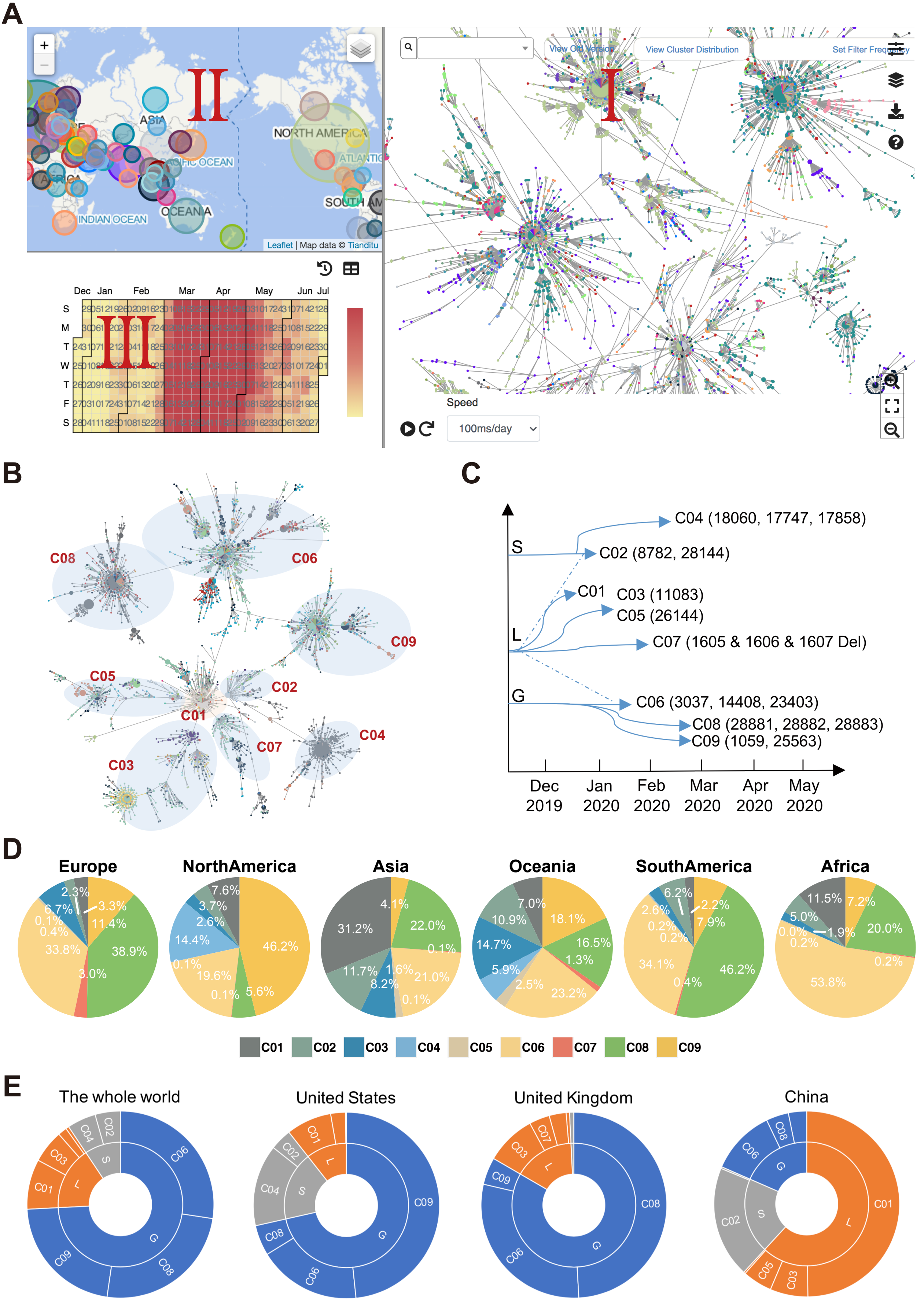

A
1
100
10k
1M
B
Partial
3730
Complete
60970
Complete high
31689, 49%
Complete Low
29281, 45%
Partial
Complete high
Complete low
C
Sequences count
10000
7500
5000
2500
0
(16,216]
(216,416]
(416,616]
(616,816]
(816,1016]
(1016,1216]
(1216,1416]
(1416,1616]
(1616,1816]
(1816,2016]
(2016,2216]
(2216,2416]
(2416,2616]
(2616,2816]
(2816,3016]
(3016,3216]
(3216,3416]
(3416,3616]
(3616,3816]
(3816,4016]
(4016,4216]
(4216,4416]
(4416,4616]
>4616
Ns count
D
Ns frequency
0.4
0.3
0.2
0.1
0.0
5'UTR
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF8
N
ORF10
3'UTR
SARS−CoV−2 genome


A
Mutation number
13000
12000
10000
8000
400
200
0
SNP
Deletion
Insertion
Indel
Mutation type
C
5'UTR, 1.71%
3'UTR, 1.68%
ORF8, 1.51%
Intergenic, 1.08%
ORF6, 0.82%
E, 0.80%
ORF10, 0.45%
ORF7a, 1.84%
M, 2.17%
ORF3a, 3.71%
S, 12.48%
ORF1ab, 66.54%
B
V367>F (Chain_C)
V367>F (Chain_A)
V367>F (Chain_B)
V367>F (Chain_C)
V367>F (Chain_A)
V367>F (Chain_B)
D
Mutation coverage
0.8
0.6
0.4
0.2
0.0
241
Ns: 0.3%
1
265
5'UTR
0.8
0.6
0.4
0.2
0.0
1059
Ns: 0.7%
3037
Ns: 0.4%
8782
Ns: 1.5%
14408
Ns: 0.2%
23403
Ns: 0.2%
266
10000
20000
29674
Genebody
0.020
0.015
0.010
0.005
0.000
29675
29903
3'UTR


A
23403
14408
3037
241
25563
28881
28882
28883
1059
28144
8782
11083
14805
18060
20268
17858
17747
26144
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2019–12–24
2020–01–08
2020–01–15
2020–01–22
2020–01–29
2020–02–05
2020–02–12
2020–02–19
2020–02–26
2020–03–04
2020–03–11
2020–03–18
2020–03–25
2020–04–01
2020–04–08
2020–04–15
2020–04–22
2020–04–29
2020–05–06
2020–05–13
2020–05–20
2020–05–27
2020–06–04
2020–06–11
2020–06–18
2020–06–24
2020–07–01
B
Virus counts
24000
20000
16000
12000
8000
4000
0
Mutation frequnecy
0.8
0.6
0.4
0.2
0
2019–12–24
2020–01–08
2020–01–15
2020–01–22
2020–01–29
2020–02–05
2020–02–12
2020–02–19
2020–02–26
2020–03–04
2020–03–11
2020–03–18
2020–03–25
2020–04–01
2020–04–08
2020–04–15
2020–04–22
2020–04–29
2020–05–06
2020–05–13
2020–05–20
2020–05–27
2020–06–04
2020–06–11
2020–06–18
2020–06–24
2020–07–01
C
Virus counts
8000
6000
4000
2000
0
Australia
Austria
Belgium
Canada
China
Denmark
France
Iceland
India
Israel
Luxembourg
Netherlands
New Zealand
Portugal
Spain
Sweden
Switzerland
Thailand
United Kingdom
United States
2019–12–24
2020–01–08
2020–01–15
2020–01–22
2020–01–29
2020–02–05
2020–02–12
2020–02–19
2020–02–26
2020–03–04
2020–03–11
2020–03–18
2020–03–25
2020–04–01
2020–04–08
2020–04–15
2020–04–22
2020–04–29
2020–05–06
2020–05–13
2020–05–20
2020–05–27
2020–06–04
2020–06–11
2020–06–18
2020–06–24
2020–07–01



1
0.8
0.6
0.4
0.2
0
Nepal(1)
Malaysia(22)
South Korea(73)
Cambodia(1)
Iran(4)
Brunei(4)
Uruguay(1)
Kenya(1)
Qatar(2)
Philippines(2)
Pakistan(8)
Indonesia(12)
Kazakhstan(53)
Singapore(442)
Kuwait(5)
China(967)
Thailand(92)
Bahrain(25)
Jamaica(8)
United Arab Emirates(32)
Japan(527)
Australia(781)
Slovenia(5)
Oman(41)
Georgia(18)
Lebanon(6)
New Zealand(173)
Norway(25)
India(1021)
United States(9341)
Tunisia(5)
Canada(367)
Saudi Arabia(124)
Chile(171)
United Kingdom(7967)
Vietnam(45)
Belgium(725)
Poland(60)
Colombia(74)
Israel(199)
Sweden(527)
Italy(157)
France(282)
South Africa(132)
Finland(234)
Bosnia and Herzegovina(31)
Peru(2)
Cyprus(8)
Mongolia(2)
Venezuela(1)
Ecuador(20)
Argentina(32)
Algeria(3)
Morocco(40)
Croatia(13)
Serbia(12)
Romania(34)
Switzerland(272)
Hungary(62)
Russia(244)
Egypt(83)
Czech Republic(36)
Luxembourg(110)
Denmark(579)
Bulgaria(15)
Bangladesh(270)
Brazil(154)
Congo(34)
Latvia(53)
Austria(361)
Portugal(413)
Senegal(49)
Gambia(6)
Netherlands(1101)
Costa Rica(6)
Germany(152)
Greece(121)
Turkey(107)
Slovakia(4)
Iceland(410)
Ireland(10)
Uganda(13)
Benin(12)
Jordan(23)
Sri Lanka(4)
Mexico(17)
Spain(593)
Nigeria(18)
2019-12-24
2020-01-08
2020-01-15
2020-01-22
2020-01-29
2020-02-05
2020-02-12
2020-02-19
2020-02-26
2020-03-04
2020-03-11
2020-03-18
2020-03-25
2020-04-01
2020-04-08
2020-04-15
2020-04-22
2020-04-29
2020-05-06
2020-05-13
2020-05-20
2020-05-27
2020-06-04
2020-06-11
2020-06-18
2020-06-24
2020-07-01


A
II
ASIA
NORTH AMERICA
PACIFIC OCEAN
ATLANTIC
SOUTH AM
AFRICA
INDIAN OCEAN
OCEANIA
Leaflet | Map data © Tianditu
View Old Version
View Cluster Distribution
Set Filter Frequency
I
III
Dec Jan Feb Mar Apr May Jun Jul
S M T W T F S
Speed
100ms/day
B
C08
C06
C09
C05
C02
C01
C07
C04
C03
C
C04 (18060, 17747, 17858)
S
C02 (8782, 28144)
C01
C03 (11083)
C05 (26144)
L
C07 (1605 & 1606 & 1607 Del)
G
C06 (3037, 14408, 23403)
C08 (28881, 28882, 28883)
C09 (1059, 25563)
Dec 2019
Jan 2020
Feb 2020
Mar 2020
Apr 2020
May 2020
D
Europe
2.3%
6.7%
0.1%
0.4%
33.8%
3.3%
11.4%
38.9%
3.0%
NorthAmerica
7.6%
3.7%
2.6%
14.4%
0.1%
19.6%
0.1%
46.2%
5.6%
Asia
4.1%
22.0%
31.2%
0.1%
11.7%
1.6%
8.2%
21.0%
0.1%
Oceania
7.0%
10.9%
18.1%
14.7%
16.5%
1.3%
5.9%
2.5%
23.2%
SouthAmerica
6.2%
2.6%
2.2%
0.2%
0.2%
7.9%
34.1%
0.4%
46.2%
Africa
11.5%
7.2%
5.0%
0.0%
1.9%
0.2%
20.0%
0.2%
53.8%
C01 C02 C03 C04 C05 C06 C07 C08 C09
E
The whole world
United States
United Kingdom
China
S
L
G
C01
C02
C03
C04
C05
C06
C07
C08
C09