

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

3 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
4 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
5 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
6 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},
7 Shuai Jiang^{1,2}, Hua Chen^{1,5}, Dali Han^{1,5}, Jingfa Xiao^{1,2,3,4}, Zhang Zhang^{1,2,3,4,*}, Wenming
8 Zhao^{1,2,3,4,*}, Yongbiao Xue^{1,2,4,*}, Yiming Bao^{1,2,3,4,*}

9
10 ¹ China National Center for Bioinformation, Beijing 100101, China

11 ² National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences,
12 Beijing 100101, China

13 ³ CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese
14 Academy of Sciences, Beijing 100101, China

15 ⁴ University of Chinese Academy of Sciences, Beijing 100049, China

16 ⁵ CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese
17 Academy of Sciences, Beijing 100101, China

18 [#]These authors contributed equally to this work.

19 ^{*}Corresponding author (baoym@big.ac.cn, ybxue@big.ac.cn, zhaowm@big.ac.cn
20 zhangzhang@big.ac.cn)

21
22 **Running title:** Song S et al /The genomic variants and haplotypes of SARS-CoV-2

23 Total word counts: 4477

24 Total References: 21

25 Total figures: 5

26 Total tables: 2

27 Total supplementary figures and tables: 3 (Figure S1, Figure S2, Table S1)

28 **Abstract**

29 On 22 January 2020, the National Genomics Data Center (NGDC), part of the China National
30 Center for Bioinformation (CNCB), created the 2019 Novel Coronavirus Resource
31 (2019nCoV), an open-access SARS-CoV-2 information resource. 2019nCoV features a
32 comprehensive integration of sequence and clinical information for all publicly available
33 SARS-CoV-2 isolates, which are manually curated with value-added annotations and quality
34 evaluated by our in-house automated pipeline. Of particular note, 2019nCoV performs
35 systematic analyses to generate a dynamic landscape of SARS-CoV-2 genomic variations at a
36 global scale. It provides all identified variants and detailed statistics for each virus isolate, and
37 congregates the quality score, functional annotation, and population frequency for each
38 variant. It also generates visualization of the spatiotemporal change for each variant and
39 yields historical viral haplotype network maps for the course of the outbreak from all
40 complete and high-quality genomes. Moreover, 2019nCoV provides a full collection of
41 SARS-CoV-2 relevant literature on COVID-19 (Coronavirus Disease 2019), including
42 published papers from PubMed as well as preprints from services such as bioRxiv and
43 medRxiv through Europe PMC. Furthermore, by linking with relevant databases in
44 CNB-NGDC, 2019nCoV offers data submission services for raw sequence reads and
45 assembled genomes, and data sharing with National Center for Biotechnology Information.
46 Collectively, all SARS-CoV-2 genome sequences, variants, haplotypes and literature are
47 updated daily to provide timely information, making 2019nCoV a valuable resource for the
48 global research community. 2019nCoV is accessible at <https://bigd.big.ac.cn/ncov/>.

49

50 **KEYWORDS:** 2019nCoV; SARS-CoV-2; Database; Genomic variation; Haplotype

51

52 **Introduction**

53 The severe respiratory disease COVID-19 [1], since its outbreak in late December 2019, has
54 rapidly spread as a pandemic. As of 14th July 2020, 12,964,809 confirmed cases have been
55 reported in 216 countries/territories/areas (WHO Situation Report Number 176;
56 <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>). As the
57 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 (GISAID) [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 (2019nCoV-2019), <https://bigd.big.ac.cn/ncov/> in CNCB-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, 2019nCoV-2019 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 14th 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 (Table 1). Here we report these significant updates of 2019nCoV-2019 and present the global landscape of SARS-CoV-2 genomes, variants and haplotypes.

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 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 2019nCoV-R by integrating all available genomes throughout the world and conducting value-added curation and analysis (**Figure 1**). As of 14th July 2020, 2019nCoV-R 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 (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 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 (6770, 50.4%) are nonsynonymous, causing amino acid changes. To gain the functional effects of those missense variants of S spike protein 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* (<https://bigd.big.ac.cn/ncov/variation/annotation>).

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 3037 and 14,408 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 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 8782 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, some 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 3037, 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 2019nCoV-2R (<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, 2019nCoV-2R 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

178 based on all identified variants beyond UTRs regions. As a result, 17,624 haplotypes were
 179 identified from 31,685 complete high-quality genome sequences as of 14th July 2020. Based
 180 on this, we construct a haplotype network for SARS-CoV-2 (**Figure 5**), a graphical
 181 representation of genomic variations by inferring relationships between individual genotypes,
 182 according to the principle of the shortest set of connections that link all nodes (genotypes)
 183 where the length of each connection represents the genetic distance [9]. To provide a whole
 184 picture of the pandemic transmission in a spatiotemporal manner, we visualize the
 185 SARS-CoV-2 haplotype network by sample collection date and across different
 186 countries/regions. It not only allows users to easily obtain a landscape of SARS-CoV-2
 187 haplotypes and their relationships, but also helps users to navigate a set of haplotypes for a
 188 specific country/region linking with additional associated information such as the number of
 189 genomes, sampling time and location (Figure 5A).

190

191 According to the haplotype network, we classify all genome sequences into nine major
 192 clusters (labelled as C01–C09; see Methods for details) (Figure 5B, 5C; **Table 2**). As the
 193 ongoing pandemic spread of SARS-CoV-2, new branches that evolve and spread faster are
 194 gradually emerging, such as clusters C04, C06, C08, and C09 (Table 2). The dominant
 195 clusters are C06 (8681, 27.4%), C08 (7,889, 24.9%), and C09 (6,940, 21.9%) (Figure 5D),
 196 which are characterized by those signature mutations of C241T, C3037T, C14408T, and
 197 A23403G, and are defined as the G clade (as the mutation at position 23,403 provoking an
 198 amino acid change D614G of S protein). These sequences have spread to 82 countries
 199 worldwide, and become the main epidemic virus type in most countries in Europe, North
 200 America, South America, Africa and West Asia, etc. For example, there were about 6827
 201 (71.5%), 8305 (83.4%), and 970 (18.5%) sequences originated from the G clade in the United
 202 States, United Kingdom, and China, respectively (Figure 5E). The wide spread and
 203 prevalence of this clade in different countries may suggest the adaptability of the virus type to
 204 human [8].

205

206 **Implementation**

207 2019nCoVVR was built based on B/S (Browser/Server) architecture. In the browser-side, it was

208 developed by JSP (Java Server Pages), HTML, CSS, AJAX (Asynchronous JavaScript and
209 XML), JQuery (a cross-platform and feature-rich JavaScript library; <http://jquery.com>) as
210 well as Semantic-UI (an open source web development framework; <https://semantic-ui.com>).
211 In the server-side, it was implemented by using Spring Boot (a rapid application development
212 framework based on Spring; <https://spring.io>). For data storage, MySQL (<https://mysql.com>)
213 was used. For interactive visualization, HighCharts (a modern SVG-based, multi-platform
214 charting library; <https://highcharts.com>), D3.js (a JavaScript library for manipulating
215 documents based on data; <https://d3js.org>) and 3Dmol.js (a JavaScript library for visualizing
216 protein structure associated with mutated amino acids) [21] were employed in 2019nCoV-2.
217 The haplotype network was implemented by d3js, Leaflet (<http://leafletjs.com>), and Echarts
218 (<http://echarts.baidu.com/>).

219

220 Discussion

221 Genome sequencing is vital to understand the epidemiology of SARS-CoV-2, since it is not
222 only useful for deciphering its genome sequences and investigating its evolution and
223 transmission, but also highly effective at determining whether individuals are part of the same
224 transmission chain [10]. According to 2019nCoV-2, however, the ratio of sequenced samples
225 to the number of confirmed cases is very low in some countries/regions (Figure S1), and even
226 genome sequences are unavailable in some affected countries/regions. The SARS-CoV-2
227 sampling bias and depth may lead to inaccurate transmission patterns and phylogenetic
228 relationships [11]. Based on sequencing all infected cases in a single region, it has proved that
229 the transmission of *Clostridium difficile* from symptomatic patients accounts for only one
230 third of all infected cases [12]. As our current understanding is still very limited, we call for
231 more efforts and collaborations in sequencing more SARS-CoV-2 genomes from both
232 symptomatic and asymptomatic cases.

233

234 Besides, as these released SARS-CoV-2 genome sequences were generated by multiple
235 different laboratories on different sequencing platforms, the quality of genome sequences is
236 another important factor, such as the Ns of genome, which may affect variant calling and
237 biased population frequency estimation. As mentioned in results, the frequency of Ns in some

238 genomic regions is high, possibly due to the low sequencing coverage, low-complexity of
 239 sequence, low-efficiency of PCR primers used in sequencing library construction, secondary
 240 structure of RNA, etc. However, most of the sequencing coverage information is unavailable,
 241 making it challenging to evaluate whether the Ns were due to low sequencing coverage. By
 242 further investigating those genomic regions with high frequency of Ns, we found (1) their GC
 243 and AG contents are close to the average of the whole genome, excluding the possibility of
 244 low complexity of sequence; (2) the length of these regions ranges from 210 to 320 bp
 245 (similar to the length of PCR product) and more than 60% of the related sequences are
 246 generated on Illumina platform (based on PCR amplification), suggesting that those Ns
 247 regions may be resulted from low-efficiency of PCR Primers during sequencing library
 248 construction; (3) by analyzing the secondary structure of these regions' RNA sequences, we
 249 found the minimum free energy is lower than those randomly extracted regions, indicating
 250 that the secondary structure is more stable and may affect the determination of genome
 251 sequences(Figure S2). Our future efforts are to construct a recognized benchmark for quality
 252 assessment and data filtration.

253

254 Compared to the early overly simplified L-S classification [8] and those comprehensive
 255 lineages defined by Rambaut et al. [13], our classification scheme with nine clusters provides
 256 a moderate system that can be correlated with the others (Table 2). The nine clusters could
 257 also be grouped into three clades defined in [8, 10], namely, S (C02 and C04), G (C06, C08
 258 and C09), and L (the rest clusters). Although haplotype network cannot give a precise
 259 evolutionary position as phylogenetic trees do, it can be used to quickly inform the clustering
 260 of viruses according to signature mutations in each haplotype. Definitely, new clusters will be
 261 introduced as the virus is continuing to evolve.

262

263 A data-driven response to SARS-CoV-2 requires a public, free, and open-access data resource
 264 that contains complete high-quality genome sequence data, and equips with automated online
 265 pipelines to rapidly analyze genome sequences. Thus, 2019nCoV (together with other
 266 resources in CNCB-NGDC) provides a wide range of data services, involving raw sequencing
 267 data archive, genome sequence and meta information management with quality control and

curation, variation analysis and data presentation and visualization. Additionally, to facilitate worldwide users to monitor any variant that may be associated with rapid transmission and high virulence, 2019nCoV-R, when compared to GISAID and NCBI Virus, features spatiotemporal dynamic tracking for all identified variants. To better understand the epidemiology of SARS-CoV-2, future directions are to collect ever 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.

276

277 **Methods**

278 **Data collection and integration**

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

283 **Quality control and curation**

284 To determine the integrity of genome sequences, one sequence is defined as ‘Complete’ if it is
285 longer than 29000 bases and covers all protein-coding/CDS regions of SARS-CoV-2 (bases
286 266:29674 of GenBank: MN908947.3); otherwise, it is defined as “Partial”. Furthermore, to
287 examine the quality of genome sequences, unknown bases (Ns) and degenerate bases (Ds,
288 more than one possible base at a particular position and sometimes referred as “mixed bases”)
289 were counted for each sequence. By our default definition, one sequence is “high-quality” if
290 $Ns \leq 15$ and $Ds \leq 50$, and “low-quality” otherwise. Besides, any sequence is clearly labelled
291 when the number of variants ≥ 15 or the total number of deletion ≥ 2 or the distribution of
292 sequence variation is more aggregated (the ratio of the number of variants divided by the total
293 number of bases in a window ≥ 0.25).

294 **Variant identification and haplotype network construction**

295 Only complete and high-quality genome sequences were used for downstream analyses,
296 including sequence comparison, variant identification, functional annotation, and haplotype

297 network construction. Genome sequence alignment was performed with Muscle (3.8.31) [19]
 298 by comparing against the earliest released SARS-CoV-2 genome (MN908947.3). Sequence
 299 variation was identified directly using an in-house Perl program. The effect of variants was
 300 determined using VEP (ENSEMBL Variant Effect Predictor) [20].

301

302 SARS-CoV-2 haplotypes were constructed based on short pseudo sequences that consist of all
 303 variants (filtering out variations located in UTR regions) only. Then, all these pseudo
 304 sequences were clustered into groups, and each group (a haplotype) represents a unique
 305 sequence pattern. The haplotype network was inferred from all identified haplotypes, where
 306 the reference sequence haplotype was set as the starting node, and its relationship with other
 307 haplotypes was determined according to the inheritance of mutations. As a result, nine major
 308 haplotype network clusters (denoted as C01–C09) were obtained according to the
 309 phylogenetic tree-and-branch structure and those shared landmark mutations (Table 1).

310 Specifically, mutations with PMF $\geq$ 5% (except for ATG deletion at position1605, PMF $\approx$ 3%)
 311 were selected, and those co-occurred mutations were determined by LD linkage analysis. A
 312 cluster was referred to sequences with those co-occurred landmark mutations.

313

314 **Data availability**

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

317

318 **CRedit author statement**

319 **Shuhui Song:** Conceptualization, Methodology, Data Analysis, Writing Original Paper,
 320 Reviewing and Editing **Lina Ma:** Data curation, Methodology, Writing **Dong Zou:** System
 321 Development, Writing **Dongmei Tian:** Methodology, Data Analysis **Cuiping Li:**
 322 Methodology, Data Analysis **Junwei Zhu:** System Development **MeiliChen:** Data curation
 323 **Anke Wang:** System Development **Yingke Ma:** System Development **MengWei Li:**
 324 Methodology, System Development **Xufei Teng:** Visualization **Ying Cui:** Data curation
 325 **Guangya Duan:** Data curation **Mochen Zhang:** Data curation **Tong Jin:** Data curation
 326 **Chengmin Shi:** Methodology **Zhenglin Du:** Methodology **Yadong Zhang:** Methodology

327 **Chuandong Liu:** Methodology **Rujiao Li:** Data curation **Jingyao Zeng:** Data curation **Lili**
 328 **Hao:** Data curation **Shuai Jiang:** Methodology **Hua Chen:** Supervision **Dali Han:**
 329 Supervision **Jingfa Xiao:** Supervision, Methodology **Zhang Zhang:** Conceptualization,
 330 Supervision, Reviewing and Editing **Wenming Zhao:** Conceptualization, Supervision,
 331 Methodology **Yongbiao Xue:** Conceptualization, Supervision **Yimin Bao:** Conceptualization,
 332 Supervision, Reviewing and Editing

333

334 **Competing interests**

335 The authors have declared no competing interests.

336

337 **Funding**

338 This work was supported by grants from The Strategic Priority Research Program of the
 339 Chinese Academy of Sciences [XDA19090116 to S.S., XDA19050302 to Z.Z.,
 340 XDB38030400 to L.M.], National Key R&D Program of China [2020YFC0848900,
 341 2016YFE0206600, 2017YFC0907502], 13th Five-year Informatization Plan of Chinese
 342 Academy of Sciences [XXH13505-05], Genomics Data Center Construction of Chinese
 343 Academy of Sciences [XXH-13514-0202], The Professional Association of the Alliance of
 344 International Science Organizations [ANSO-PA-2020-07], The Open Biodiversity and Health
 345 Big Data Programme of IUBS, International Partnership Program of the Chinese Academy of
 346 Sciences [153F11KYSB20160008]. K. C. Wong Education Foundation to Z.Z., and The
 347 Youth Innovation Promotion Association of Chinese Academy of Science [2017141 to S.S.,
 348 2019104 to L.M.]; Funding for open access charge: The Strategic Priority Research Program
 349 of the Chinese Academy of Sciences.

350

351 **Acknowledgements**

352 We thank our colleagues and students for their hard working on the 2019nCoV
 353 (<https://bigd.big.ac.cn/ncov>). We also thank a number of users and CNCB-NGDC members
 354 for reporting bugs and sending comments. Complete genome sequences used for analyses
 355 were obtained from the CNCB-NGDC, CNGBdb, GenBank, GISAID, and NMDC databases.
 356 We acknowledge the sample providers and data submitters listed on Table S1

357

358 **Authors' ORCID ID**

359 0000-0003-2409-8770 (Shuhui Song)

360 0000-0001-6390-6289 (Lina Ma)

361 0000-0002-7169-4965 (Dong Zou)

362 0000-0003-0564-625X (Dongmei Tian)

363 0000-0002-7144-7745 (Cuiping Li)

364 0000-0003-4689-3513 (Junwei Zhu)

365 0000-0003-0102-0292 (Meili Chen)

366 0000-0002-2565-2334 (Anke Wang)

367 0000-0002-9460-4117 (Yingke Ma)

368 0000-0001-6163-2827 (Mengwei Li)

369 0000-0001-9282-4282 (Xufei Teng)

370 0000-0001-9201-0465 (Ying Cui)

371 0000-0003-4582-5156 (Guangya Duan)

372 0000-0001-9136-451X (MoChen Zhang)

373 0000-0003-0791-2822 (Tong Jin)

374 0000-0003-0237-4092 (Chengmin Shi)

375 0000-0003-2147-3475 (Zhenglin Du)

376 0000-0003-0918-5673 (Yadong Zhang)

377 0000-0002-9904-7786 (Chuandong Liu)

378 0000-0002-3276-8335 (Rujiao Li)

379 0000-0001-7364-9677 (Jingyao Zeng)

380 0000-0003-3432-7151 (Lili Hao)

381 0000-0002-6722-176X (Shuai Jiang)

382 0000-0002-9829-6561 (Hua Chen)

383 0000-0001-7119-1578 (Dali Han)

384 0000-0002-2835-4340 (Jingfa Xiao)

385 0000-0001-6603-5060 (Zhang Zhang)

386 0000-0002-4396-8287 (Wenming Zhao)

387 0000-0002-6895-8472 (Yongbiao Xue)

388 0000-0002-9922-9723 (Yimin Bao)

389

390

391 **References**

- 392 [1] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species
393 Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it
394 SARS-CoV-2. *Nat Microbiol* 2020;5:536–44.
- 395 [2] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human
396 respiratory disease in China. *Nature* 2020;579:265–9.
- 397 [3] Zhang Z, Song S, Yu J, Zhao W, Xiao J, Bao Y. The Elements of Data Sharing. *Genomics*
398 *Proteomics Bioinformatics* 2020;18:1–4.
- 399 [4] Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality.
400 *Euro Surveill* 2017;22:30494.
- 401 [5] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence
402 (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic*
403 *Acids Res* 2016;44:D733–45.
- 404 [6] Zhao WM, Song SH, Chen ML, Zou D, Ma LN, Ma YK, et al. The 2019 novel coronavirus resource.
405 *Yi Chuan* 2020;42:212–21. (in Chinese with an English abstract)
- 406 [7] Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of
407 SARS-CoV-2. *National Science Review* 2020;7:1012–23.
- 408 [8] Korber B, Fischer W, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline
409 reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* 2020;
410 <https://doi.org/10.1101/2020.04.29.069054>.
- 411 [9] Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. *Mol*
412 *Biol Evol* 1999;16:37–48.
- 413 [10] Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission.
414 *Curr Opin Microbiol* 2015;23:62–7.
- 415 [11] Mavian C, Pond SK, Marini S, Magalis BR, Vandamme AM, Dellicour S, et al. Sampling bias and
416 incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. *Proc Natl*
417 *Acad Sci U S A* 2020;117:12522–3.
- 418 [12] Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, et al. Diverse sources of *C.*
419 *difficile* infection identified on whole-genome sequencing. *N Engl J Med* 2013;369:1195–205.
- 420 [13] Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature
421 proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020.
422 <https://doi.org/10.1038/s41564-020-0770-5>
- 423 [14] Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2
424 Transcriptome. *Cell* 2020;181:914–21 e10.
- 425 [15] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to
426 global health. *Glob Chall* 2017;1:33–46.
- 427 [16] National Genomics Data Center M, Partners. Database Resources of the National Genomics Data
428 Center in 2020. *Nucleic Acids Res* 2020;48:D24–D33.

- 429 [17] Shi W, Qi H, Sun Q, Fan G, Liu S, Wang J, et al. gcMeta: a Global Catalogue of Metagenomics
430 platform to support the archiving, standardization and analysis of microbiome data. *Nucleic Acids Res*
431 2019;47:D637–D48.
- 432 [18] Xiao SZ, Armit C, Edmunds S, Goodman L, Li P, Tuli MA, et al. Increased interactivity and
433 improvements to the GigaScience database, GigaDB. *Database (Oxford)* 2019;2019:1–9.
- 434 [19] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
435 *Nucleic Acids Res* 2004;32:1792–7.
- 436 [20] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect
437 Predictor. *Genome Biol* 2016;17:122.
- 438 [21] Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*
439 2015;31:1322–4.

440

441 **Figure legends**

442 **Figure 1 Statistics and distribution of all released SARS-CoV-2 genomes.** (A) Distribution
443 of released genome sequences by country, territory or region; (B) Number and percentage of
444 complete and high-quality genomes; (C) Sequences number of different Ns ranges for
445 low-quality genomes; (D) Frequency distribution of Ns across the whole genome.

446 **Figure 2 Landscape of genomic variants.** (A) Numbers of different mutation types
447 (including SNPs, deletion, insertion, and indels; the orange bar represents all mutations while
448 the blue bar represents mutations with PMF>0.001); (B) Structure display for
449 nonsynonymous mutations; (C) Pie chart of variant annotation for each gene; (D) Population
450 mutated frequency (PMF) for all variants.

451 **Figure 3 Spatiotemporal dynamics of genomic variants.**(A) Population mutation frequency
452 (PMF) of variants over time; (B) PMF and cumulative sequence growth curve for position
453 (n23403, pD614G); (C) Cumulative growth curve of the number of mutated sequences in
454 selected countries for position (n23403, pD614G).

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

456 **Figure 5 Haplotype network and cluster identification and distribution.** (A) The snapshot
457 of haplotype network dashboard, which can dynamically show the development of haplotype
458 (I) across countries (II) and over time (III). Each node in the network represents a haplotype
459 and the node size is proportional to the number of viral genome sequences, where the edge
460 between any two nodes represents the genetic distance between two haplotypes (i.e. the
461 number of mutation sites); (B-C) Schematic diagram of haplotype clusters (C01–C09) and

462 their corresponding common mutation sites for each cluster; (D) Distribution of C01–C09
463 clusters across different continents; (E) Distribution of different clusters throughout the world
464 and in three representative countries (US, UK and China).

465

Supplementary Figure and Table legend

467 Figure S1 Distribution of genome sequence count divided by the number of confirmed cases
468 for each country.

469 Figure S2 Compositional analysis of whole genome and two representative genomic regions
470 with high frequency of Ns. (A) GC and AG Compositional variability; (B) Two representative
471 genomic regions with high frequency of Ns and distribution of sequencing platforms for those
472 corresponding sequences; (C) The secondary structure of one representative Ns region and
473 one non-Ns region.

474 Table. S1 Coronavirus sequence datasets used for the study.

475

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

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 3D structures (S Protein)		✓
Spatiotemporal dynamics analysis of genomic variants		✓
Haplotype network and dynamic evolution		✓
AI diagnosis and online tools		✓
Enhanced user-friendly interface		✓

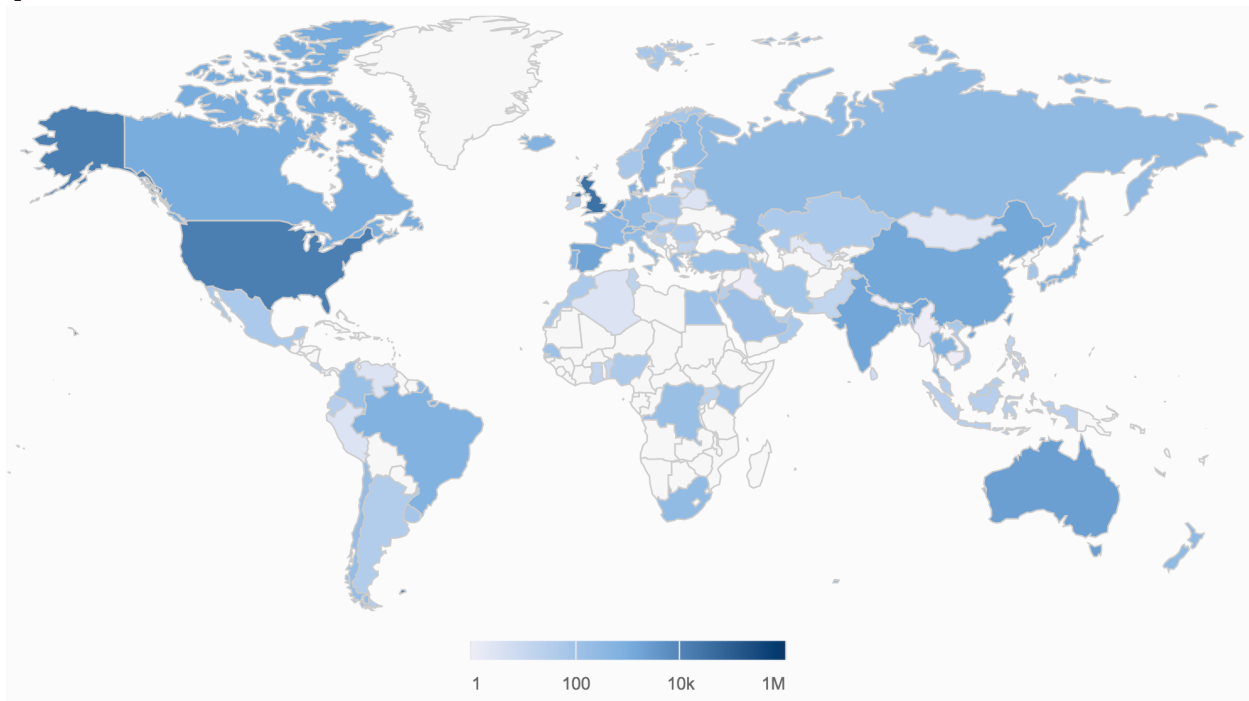
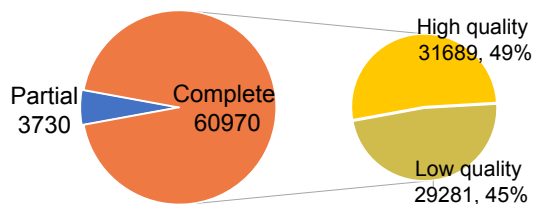
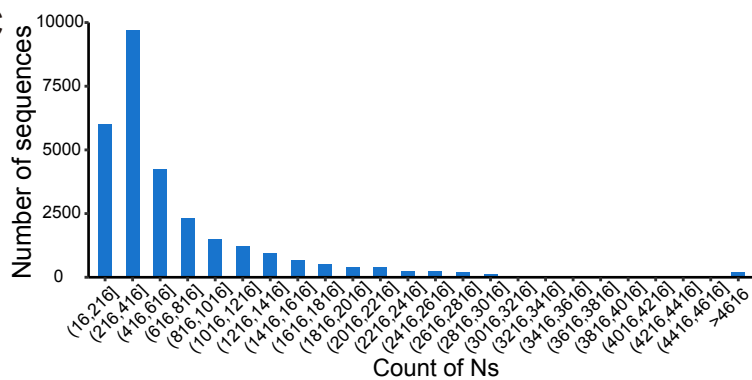
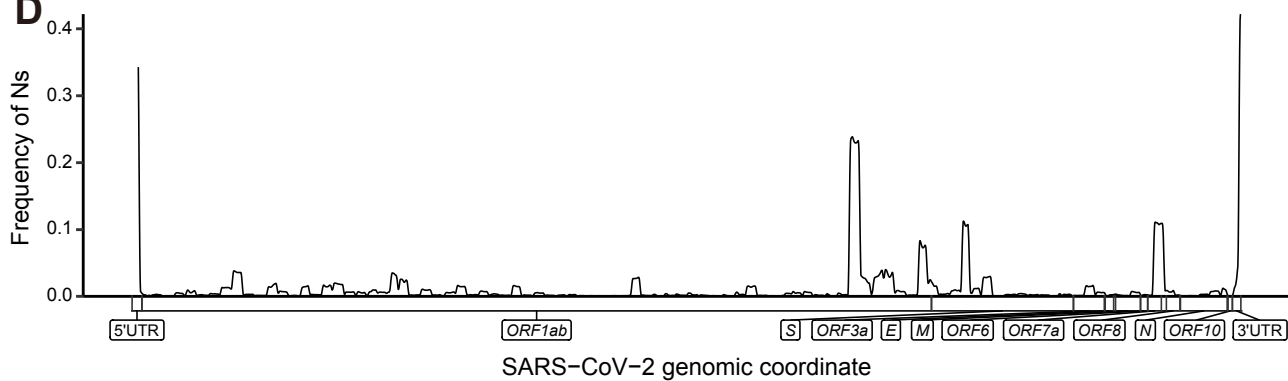
477

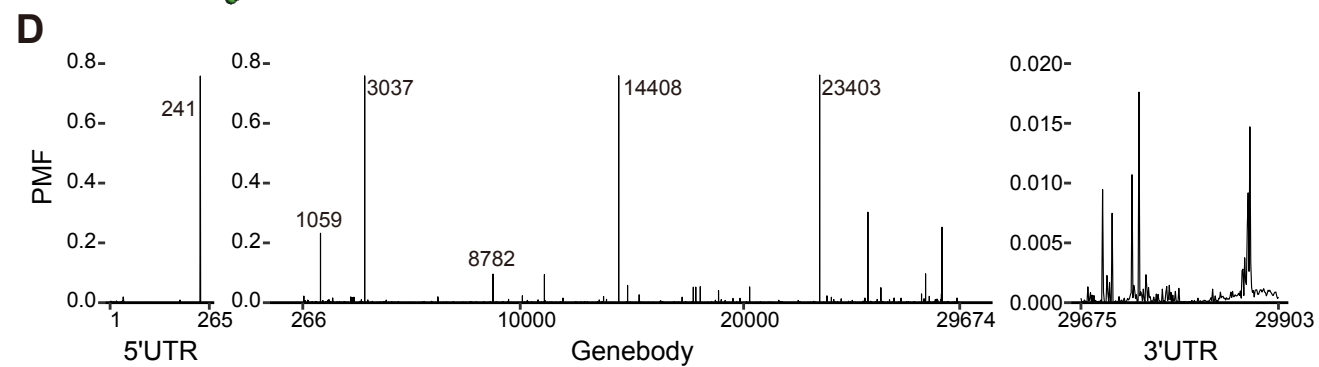
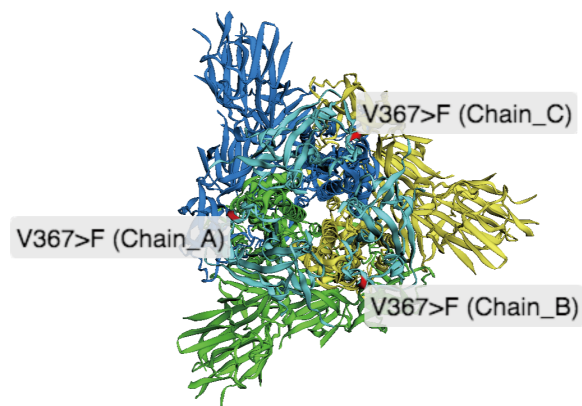
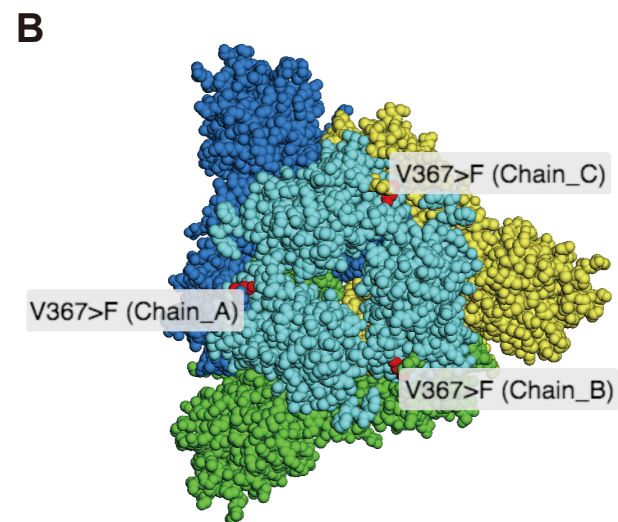
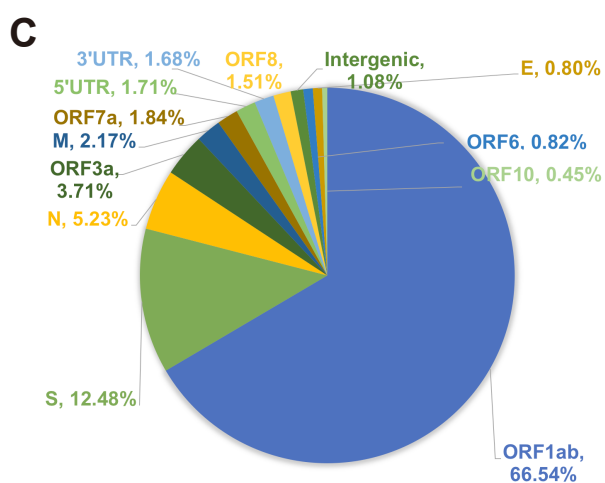
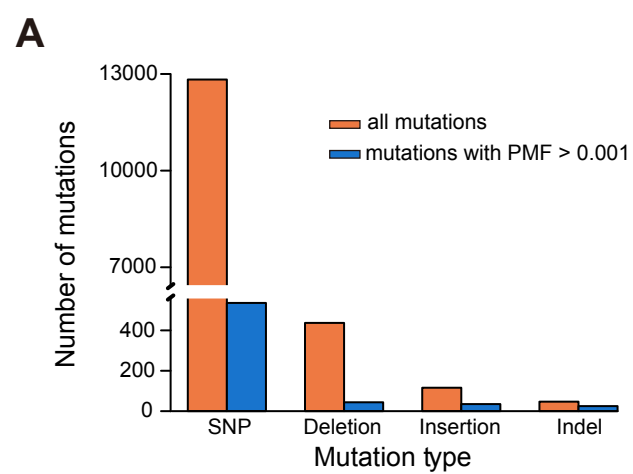
Table 2 Signature mutations of haplotype clusters

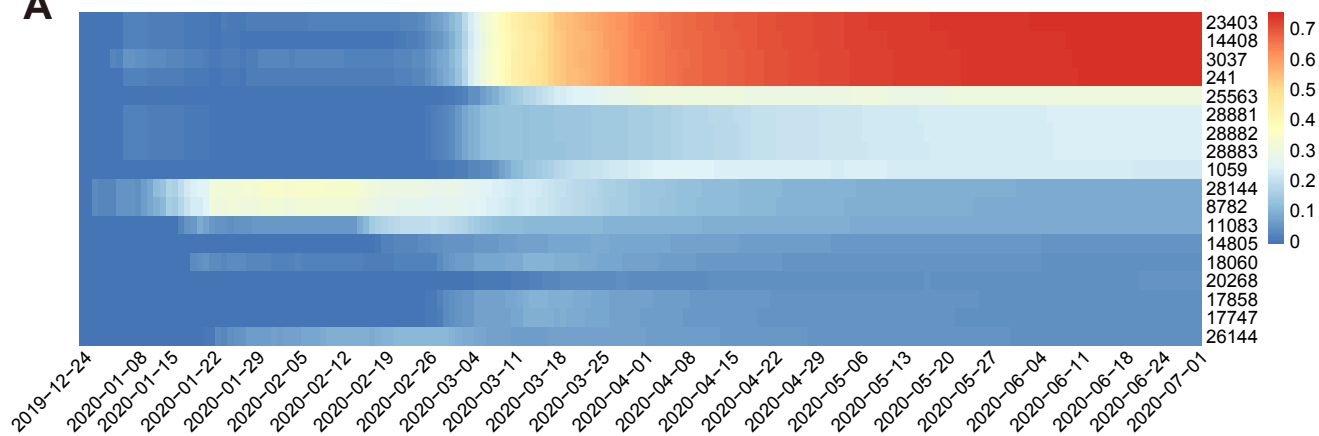
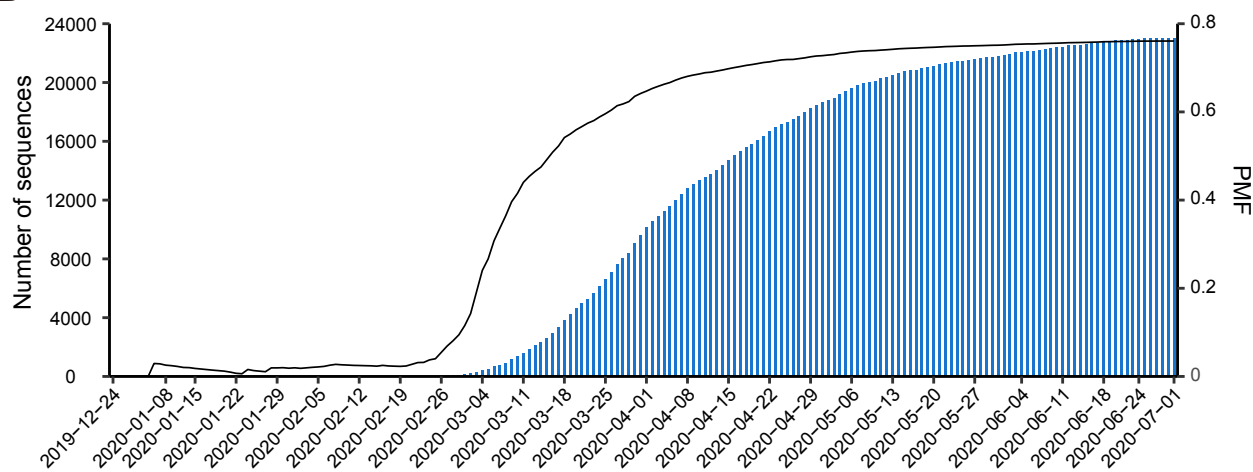
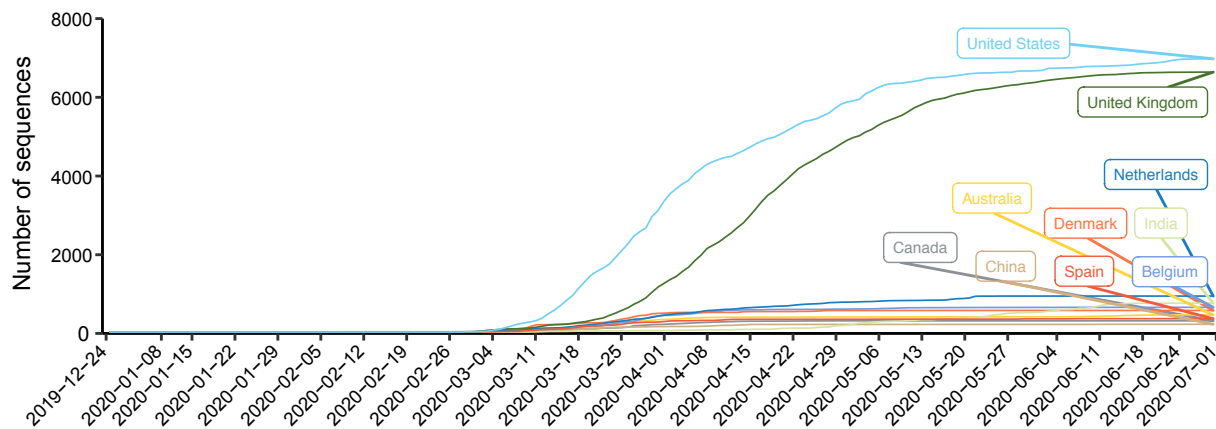
Clade	Cluster ID	Genomic Location	Gene/Genomic Regions	Mutation	Amino acid Position and Change	Mutation Frequency	Number of Mutated Sequences	Classification in Lu et.al 2020 NSR ^a	Lineage in Rambaut et.al 2020 Nat Microbiol ^b
L	C01	NA	NA	NA	NA	NA	NA	L	B/B.3/B.1.3 etc.
	C03	11083	<i>ORF1ab</i>	G->T	p.3606L>F	0.09	2982	L	B.2/B.2.1/B.4 etc.
	C05	26144	<i>ORF3a</i>	G->T	p.251G>V	0.05	1592	L	B/B.2
	C07	1604	<i>ORF1ab</i>	AATG->A	p.447-448ND>N	0.02	503	L	B/B.8
S	C02	8782	<i>ORF1ab</i>	C->T	p.2839S	0.09	3034	S	A/A.3/A.4/A.5
		28144	<i>ORF8</i>	T->C	p.84L>S	0.09	3063		
	C04	17747	<i>ORF1ab</i>	C->T	p.5828P>L	0.05	1644	S	A.1/A.1.1/A.1.2
		17858	<i>ORF1ab</i>	A->G	p.5865Y>C	0.05	1657		
		18060	<i>ORF1ab</i>	C->T	p.5932L	0.05	1695		
G	C06	241	<i>5'UTR</i>	C->T		0.75	24028	L	B.1/B.1.5/B.1.11 etc.
		3037	<i>ORF1ab/nsp4</i>	C->T	p.924F	0.75	24045		
		14408	<i>ORF1ab/RdRp</i>	C->T	p.4715P>L	0.75	24055		
		23403	<i>S</i>	A-> G	p.614D>G	0.76	24128		
	C08	28881	<i>N</i>	G->A	p.203R>K	0.25	8003	L	B.1/B.1.1/B.1.10 etc.
		28883	<i>N</i>	G-> C	p.204G>R	0.25	7995		
		28882	<i>N</i>	G->A	p.203R	0.25	7985		
	C09	1059	<i>ORF1ab</i>	C->T	p.265T>I	0.23	7357	L	B.1/B.1.21/B.1.43
		25563	<i>ORF3a</i>	G->T	p.57Q>H	0.30	9594		

Note: ^a 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.

^b 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.

A**B****C****D**



A**B****C**

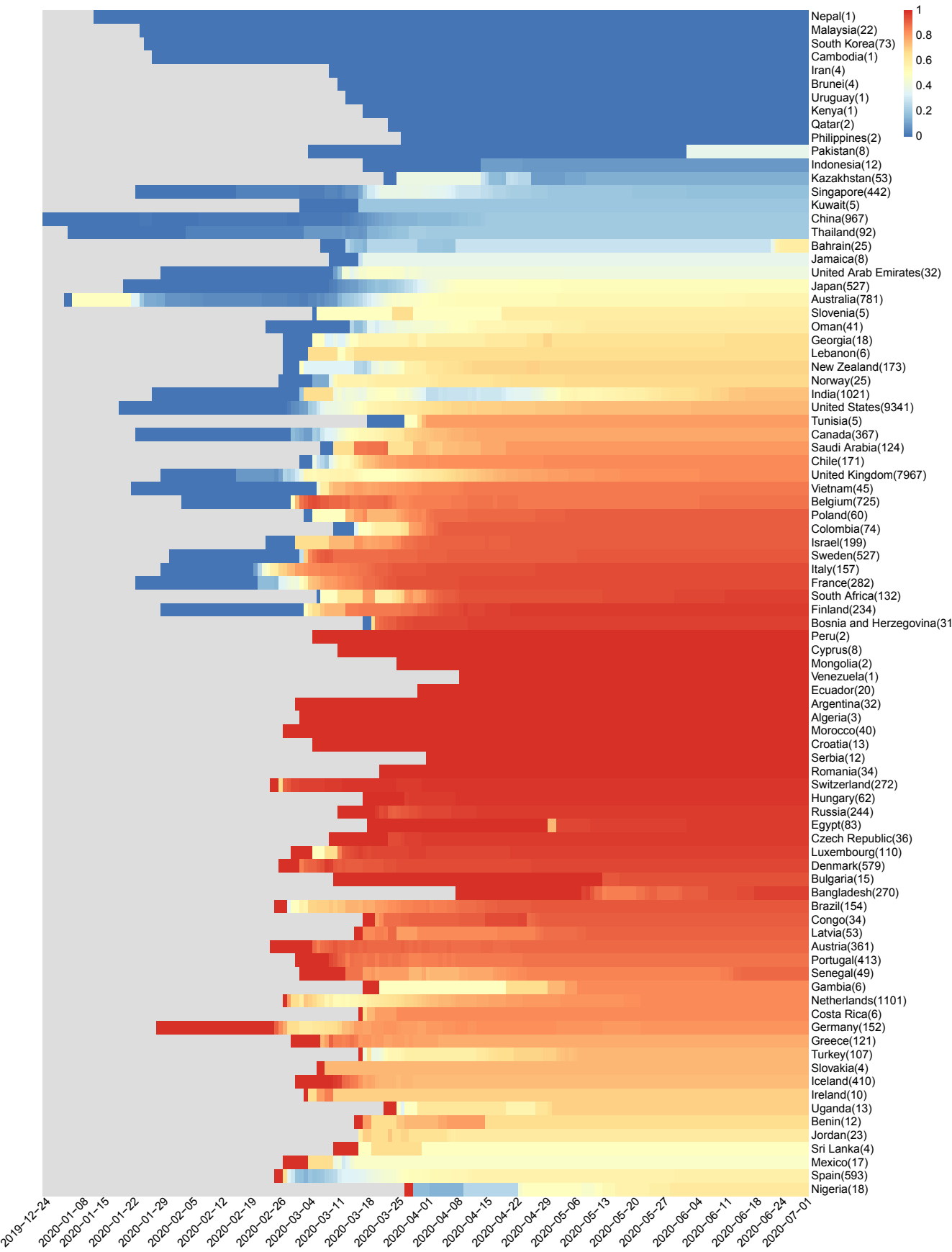


Figure 1 displays six pie charts representing the distribution of 10 categories (C01-C09) across six regions: Europe, North America, Asia, Oceania, South America, and Africa. The categories are color-coded: C01 (dark grey), C02 (teal), C03 (dark blue), C04 (light blue), C05 (tan), C06 (yellow), C07 (red), C08 (green), and C09 (orange).

Region	C01	C02	C03	C04	C05	C06	C07	C08	C09
Europe	0.1%	0.4%	6.7%	2.3%	3.3%	33.8%	3.0%	38.9%	11.4%
North America	0.1%	0.1%	19.6%	14.4%	7.6%	46.2%	5.6%	0.1%	2.6%
Asia	31.2%	8.2%	11.7%	1.6%	21.0%	0.1%	0.1%	22.0%	4.1%
Oceania	10.9%	7.0%	14.7%	5.9%	2.5%	23.2%	1.3%	16.5%	18.1%
South America	0.2%	0.2%	2.6%	6.2%	2.2%	34.1%	0.4%	46.2%	7.9%
Africa	11.5%	5.0%	0.0%	0.2%	1.9%	53.8%	0.2%	20.0%	7.2%

The figure displays four donut charts representing the distribution of COVID-19 cases by country/region and age group. The charts are labeled Global, US, UK, and China. Each chart has an outer ring for country/region and an inner ring for age group (G, L, S).

- Global:** The outer ring shows C06 (blue), C09 (blue), C08 (blue), C03 (orange), C01 (orange), C04 (orange), and C02 (grey). The inner ring shows G (blue), L (orange), and S (grey).
- US:** The outer ring shows C09 (blue), C06 (blue), C04 (grey), C03 (orange), C02 (grey), and C01 (orange). The inner ring shows G (blue), L (orange), and S (grey).
- UK:** The outer ring shows C08 (blue), C06 (blue), C09 (blue), C01 (orange), C07 (orange), C03 (orange), and C09 (blue). The inner ring shows G (blue), L (orange), and S (grey).
- China:** The outer ring shows C03 (orange), C01 (orange), C05 (orange), C02 (grey), C05 (orange), and C01 (orange). The inner ring shows G (blue), L (orange), and S (grey).