

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38

Supplemental methods and materials

Contents

Data quality control.....	2
Distributed genome alignments.....	3
Ancestral alleles of SARS-CoV-2.....	4
Construction of the evolutionary tree based on distributed alignments	4
Imputation of ambiguous and missing nucleotides	5
Parsimony inference of mutations for strains in each branch	6
Mutations affected by recombination.....	6
Mutation cold spots.....	7
Estimation of mutation rate.....	9
Maximum-likelihood phylodynamic analysis.....	11
Displaying SARS-CoV-2 genomic variants in tree-based file format.....	12
Maximum-likelihood analysis based on the existing mutation-annotated tree	13
CGB binary nomenclature for each internal node or branch.....	13
Visualization of the huge evolutionary tree by movie-maker strategy.....	14
Data searching, filtering, and visualization of a single clade on the huge tree	15
Tree visualization with CGB.....	15
Coordinated annotation tracks	15
Detection of branch-specific accelerated evolution of SARS-CoV-2	18
Sequential occurrence of B.1.1.7-associated mutations.....	19
No correlation between ORF1ab and spike gene mutation rates	19
Detection of SARS-CoV-2 strains that evolve more slowly recently	20
Detection of on-going selection of SARS-CoV-2	21
Lineage tracing.....	25
Analysis of SARS-CoV-2 transmission in Washington State.....	26
Phylogenetic analysis of strains in the Auckland outbreak	27
Acknowledgement of the person who first discovers any SARS-CoV-2 variant.....	28
D614G mutation in the spike protein	28
Standalone and web-based CGB.....	29
CGB in multiple languages	29
Timely update of CGB	30
Statistical information.....	30
Data availability	30
References.....	30

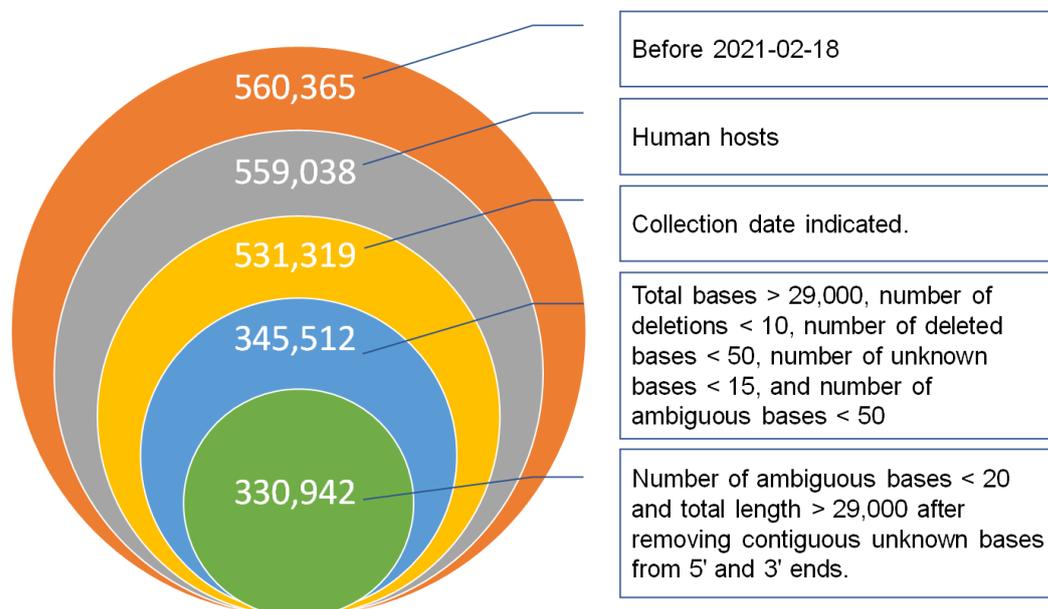
1

Supplemental methods and materials

2 **Data quality control**

3 SARS-CoV-2 genomic sequences were obtained from the 2019nCoV database (1)
4 established by China National Center for Bioinformation (CNCB). Detailed
5 information on this database is available at
6 https://bigd.big.ac.cn/ncov/release_genome. All SARS-CoV-2 isolates are from
7 humans. To obtain high-quality SARS-CoV-2 genomic sequences, quality control
8 measures were applied (Figure S1).

9



10

11 **Figure S1. Quality control pipeline.** The value in each circle is number of sequences
12 identified in the quality control performed on 18 February, 2021.

13

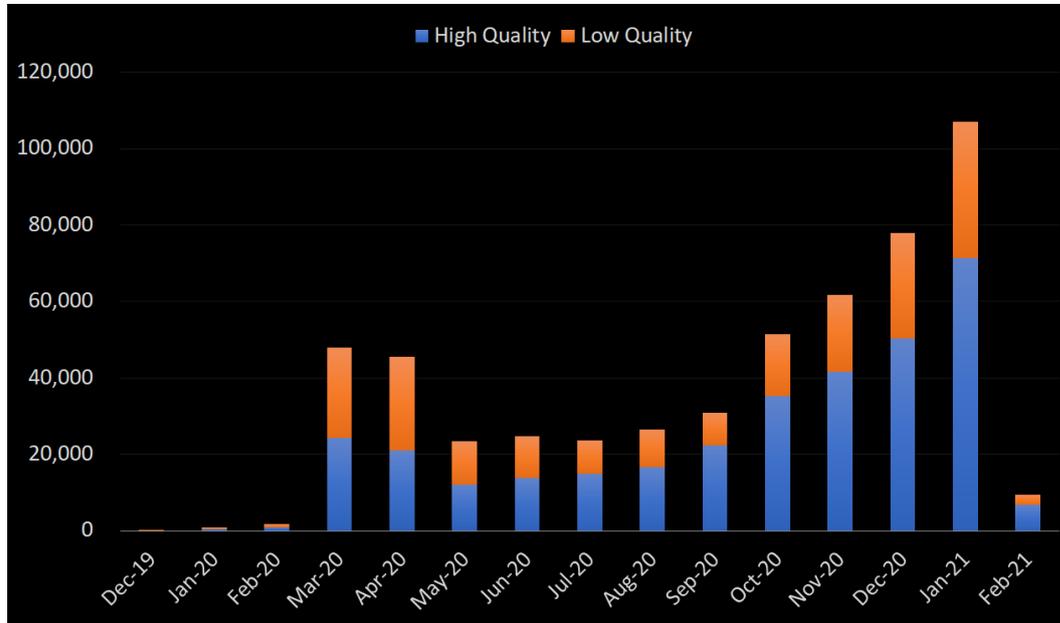
14

15 The following criteria were used to select high quality sequences. First, the collection
16 date of each strain is indicated. Second, the sequence length is longer than 29,000
17 bases, and the genome contains all protein-coding genes. Third, a gap found by
18 sequence alignment is considered as one deletion, the number of deletions is <10, and
19 the number of deleted bases is < 50. Forth, the number of unknown bases (Ns) is < 15,
20 and the number of ambiguous bases (Ds) is < 50. Fifth, the length of the genome is
21 longer than 29,000 bases after removing contiguous unknown bases from 5' and 3'
22 ends. Sixth, as analysis of 23,336 genomes revealed that 5% of the genomes contain
23 more than 19 ambiguous (Ds) and unknown (Ns) bases, a high-quality sequence must
24 have a total number of ambiguous and unknown bases < 20.

25

1 After applying these criteria, 330,942 high-quality genomic sequences were identified
2 and used for subsequent analyses, unless noted otherwise. The number of identified
3 high- and low-quality genomes in each month is shown in Figure S2.

4
5



6

7 **Figure S2. Number of high- and low-quality SARS-CoV-2 genomic sequences at**
8 **various time points.**

9 **Distributed genome alignments**

10 Genome alignment was performed using the software MAFFT (2) with parameters
11 “--auto --addfragments” after dividing input sequences into reference (GenBank
12 accession number: NC_045512) (3) and others. Because of the explosion in
13 SARS-CoV-2 genomic data, it is nearly impossible to perform daily update with the
14 currently available analysis framework. To solve this problem, the distributed
15 alignment system was developed (Figure 1), which reduces the total alignment time
16 complexity to $\mathcal{O}(n)$, where $\mathcal{O}(\cdot)$ is a linear function, and n is number of viral
17 strains. In this study, each alignment contained approximately 5,000 genomic
18 sequences, including the reference SARS-CoV-2 sequence (NC_045512) (3). To
19 generate the outgroup alignment file, the reference sequence (NC_045512) (3) was
20 aligned with the sequences of two outgroups: bat coronavirus RaTG13 (4) and
21 pangolin coronavirus PCoV-GX-P1E (5).

1 **Ancestral alleles of SARS-CoV-2**

2 In total, 272 SARS-CoV-2 strains were collected before 31 January, 2020. These
3 strains were collectively named “early samples” in this study. To detect ancestral
4 alleles, the region between nucleotide positions 100 and 29,800 of each genome was
5 examined. Compared to the reference sequence (NC_045512) (3), 28,846
6 monomorphic and 855 polymorphic sites were detected in the genomes of early
7 samples, and the ancestral alleles for those sites are determined. Upon further
8 comparison with the sequences of the two outgroups (RaTG13 and PCoV-GX-P1E)
9 (4, 5), the majority of major alleles in 827 (96.7%) of the 855 polymorphic sites were
10 found to be identical to the alleles in the outgroup genomes. Among the 28 unique
11 polymorphic sites, minor alleles in 26 sites were found to be rare with a frequency
12 less than 0.06, suggesting that the major alleles in these 26 sites in the early samples
13 are ancestral. The frequencies of two major alleles *8,782C* and *28,144T* are 0.684 and
14 0.640, respectively. The minor alleles are *8,782T* and *28,144C*. Examination of seven
15 SARS-CoV-2 strains collected in December 2019 revealed that they all carry these
16 two major alleles, suggesting that they are ancestral alleles. On the evolutionary tree,
17 the most recent common ancestor (MRCA) of SARS-CoV-2 is located at the root of
18 the tree and found to harbor all of these ancestral alleles. The sequence between
19 nucleotide position 100 and 29,800 of MRCA was found to be identical to that of the
20 reference genome sequence (GenBank accession number: NC_045512) (3). The
21 finding is consistent with that of a previous study (6).

22 **Construction of the evolutionary tree based on distributed alignments**

23 To build the evolutionary tree, the sequence corresponding to the reference sequence
24 between nucleotides 100 and 29,800 of each genome was used. Initially, the tree was
25 built using the software FastTree (7) and a slightly revised version of RAxML (8). To
26 accommodate the entire length of each SARS-CoV-2 genome, the minimum branch
27 length was changed from 10^{-5} to 10^{-10} in RAxML. However, these two methods
28 were later found to be unsatisfactory because both FastTree and RAxML cannot
29 analyze distributed alignments and sub-genomic regions. Furthermore, to use
30 FastTree and RAxML, a unified multiple sequence alignment must be done for daily
31 updates. This is beyond the capability of our computing facility. FastTree and
32 RAxML also cannot distinguish missing bases from indels because both appear as “-”
33 in the alignments. As gaps are ignored by these two methods and indels provide
34 valuable information for construction of phylogenetic tree of closely related
35 SARS-CoV-2 strains, new approaches are needed to accomplish the task. To simplify
36 CGB implementation, the Neighbor-Joining method (9) was used.

37

1 When calculating genetic distances, five different features are considered. First,
2 missing bases at 5' and 3' ends (presented as gaps in alignments) are ignored. Second,
3 insertions and deletions are taken into consideration. Third, IUPAC (International
4 Union of Pure and Applied Chemistry) ambiguous nucleotide characters (e.g., Y and
5 R) are supported. As disambiguating nucleotides will generate a huge number of
6 artificial sequences, genetic distances would be overestimated if all possible
7 sequences are compared.

8

9 To solve this problem, the following strategy was used to treat ambiguous bases. For
10 comparison of the sequence ACGRRCG with the reference sequence ACGRACG,
11 ACGRRCG is converted to ACGRACG and ACGRGCG. The resulting 2 new sequences
12 are defined as one sequence set. Because this sequence set has the sequence
13 ACGRACG that is the same as that of the reference sequence, the strain with the
14 sequence ACGRRCG is considered as the same type as the strain with the reference
15 sequence ACGRACG. For comparison of the sequence ACGRRCG with the sequence
16 ACGRYCG, ACGRRCG is converted to ACGRACG and ACGRGCG, and ACGRYCG is
17 converted to ACGRCCG and ACGRTCG. Therefore, two sequence sets are generated.
18 Because the four sequences in these two sequence sets are different, the strain with
19 the sequence ACGRRCG and the one with the sequence ACGRYCG are considered as
20 two different types. For comparison of the sequence ACGRRCG with the sequence
21 ACGRHCG, ACGRRCG is converted to ACGRACG and ACGRGCG, and ACGRHCG is
22 converted to ACGRACG, ACGRCCG and ACGRTCG. As the resulting two sequence sets
23 share the same sequence ACGRACG, the strain with the sequence ACGRRCG and the
24 one with the sequence ACGRHCG are considered as the same type.

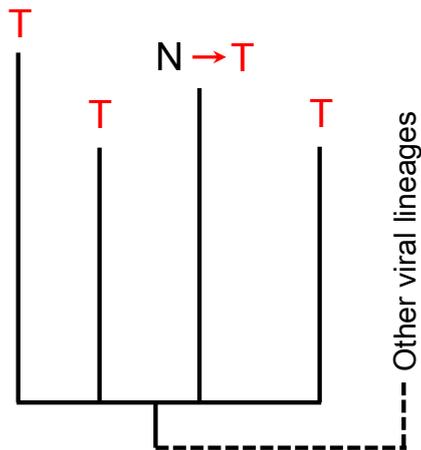
25

26 Forth, the sequences of two genomes for comparison are placed in different
27 alignments, and the sequence of the reference genome is used as the coordinate for
28 nucleotide positions. Fifth, the genetic distance between outgroups and a
29 SARS-CoV-2 strain is determined after adding two components: the average genetic
30 distance between outgroups and the most recent common ancestor (MRCA), and the
31 genetic distance between MRCA and the strain.

32 **Imputation of ambiguous and missing nucleotides**

33 An ambiguous or missing base can be imputed (Figure S3) if the strain with the
34 ambiguous base shares the same phylogeny with neighboring lineages (10). For this
35 imputation, the allele frequency and the definition of IUPAC ambiguous nucleotide
36 characters are considered, and only the lineages with collection dates ± 30 days apart
37 are compared.

38



1
 2 **Figure S3. Imputation of ambiguous nucleotides of a lineage using the**
 3 **information of its siblings.**

4 **Parsimony inference of mutations for strains in each branch**

5 After ambiguous and missing nucleotides are replaced with inferred nucleotides,
 6 mutations in strains of each branch are recapitulated according to the principle of
 7 parsimony (11, 12). Although the analysis is performed site by site, large deletions
 8 spanning over a number of regions are merged as a single large deletion, and a long
 9 insertion is considered as a united element. Thus it is easy to trace recurrent deletions
 10 (13) whenever necessary.

11 **Mutations affected by recombination**

12 To determine the effect of recombination on evolution, it is necessary to understand the
 13 history of recombination which is usually represented by the ancestral recombination
 14 graph (ARG) (14-16). Because it is impossible to construct an ARG for the huge
 15 collection of SARS-CoV-2 variants, a new method needs to be developed. According
 16 to the finite sites model, which is commonly used to study fast evolving organisms
 17 (17), recombination and recurrent mutation can generate similar genomic variants
 18 (Figure S4). As recombination creates a hybrid genomic structure (18), it can be
 19 distinguished from a recurrent mutation (Figure S4), which affects only the mutated
 20 site. In contrast, a recombination event affects a large part of genome (Figure S4A).

21
 22

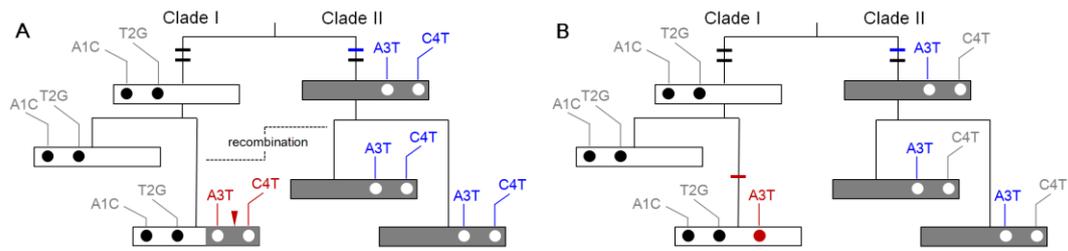


Figure S4. Generation of similar genomic variants by recombination and recurrent mutation.

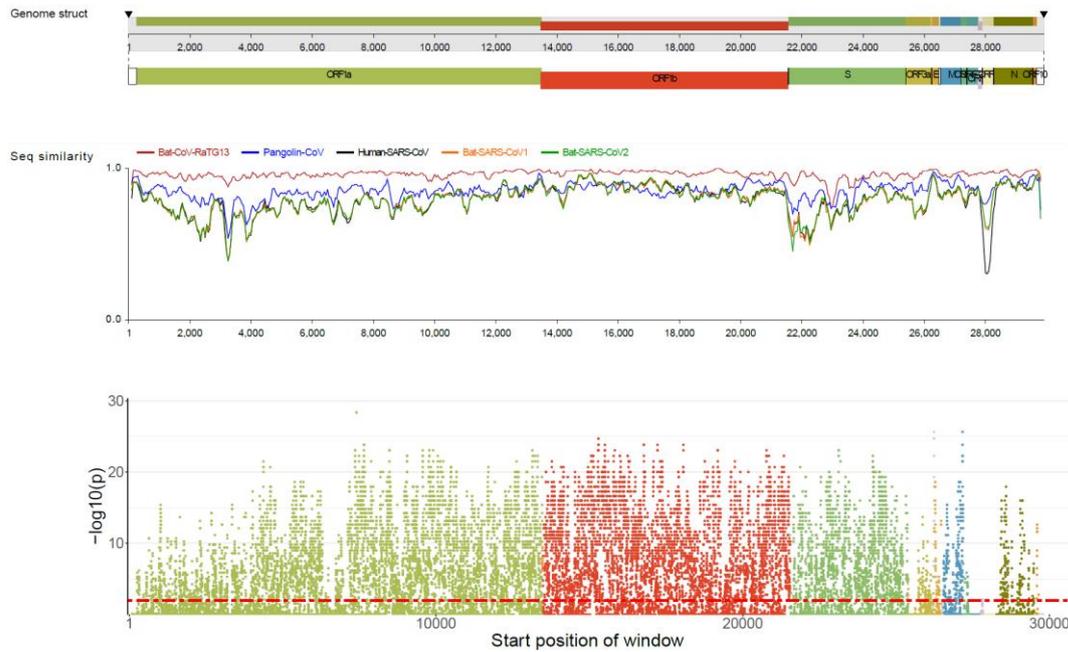
A) Recombination creates a hybrid genomic structure. The region affected by recombination is indicated with a red arrowhead. Each notch of the branches represents a mutation. Open and dark gray square strips represent sequences in two lineages. Solid and empty circles denote mutations. In clade I, mutations A3T and C4T are observed due to recombination. These two mutations are considered recurrent if the recombination is ignored because they are also present in clade II.

B) A recurrent mutation A3T, marked in red, occurs in clade I. The same mutation (marked in blue) also occurs in clade II.

To identify mutations due to recombination, a flagging procedure is performed in four steps. First, multiple mutations that occur at the same genomic position, all mutations are labeled with a recombination flag. Second, mutations are categorized according to their types. Different mutations are considered as the same type if their ancestral and derived alleles are the same. Third, for each category, the recombination flag of the most prevalent mutation is removed because this mutation is unlikely caused by recombination. The prevalence of a certain mutation corresponds to the number of its descendants (19). Back mutations are not considered. Forth, if two recombination-flagged mutations are less than 20 kb apart, their recombination flags are maintained.

Mutation cold spots

An analysis with a 10-base sliding window and a sliding step of one base was performed to identify mutation cold spots (Figure S5), which are areas in the genome with mutation rates lower than the average mutation rate of the entire genome. To avoid the effect of recombination on the determination of mutation cold spots, recombination-flagged mutations were excluded.



1 **Figure S5. Manhattan plot of mutation cold spots in the genome of SARS-CoV-2.**
 2 Results of genome-wide scan for mutation cold spots are shown in Manhattan plot of
 3 significance against SARS-CoV-2 reference genomic locations. In total, 330,942 high
 4 quality genomic sequences (submitted before February 18, 2021) were analyzed. Each
 5 dot represents one window. P -values are FDR-corrected. The dotted red line denotes
 6 FDR-corrected P -value < 0.01 . Dots above the line represent mutation cold spots.
 7 Genomic structure and sequence similarity between SARS-CoV-2 reference genome
 8 (NC_045512.2) (3) and the genomes of five other coronaviruses are shown above the
 9 Manhattan plot.
 10

11
 12 To find mutation cold spots, the mutation density of a genome is denoted as β
 13 (mutations per base), and the observed number of mutations within a 10-base window
 14 is denoted as ξ_{obs} . Under the assumption of homogeneous mutation distribution, the
 15 expected number of mutations within the window is 10β . The significant level of
 16 mutation cold spots is determined by Poisson probability (20, 21): $P(x \leq \xi_{obs}) =$

17 $\sum_{x \leq \xi_{obs}} e^{-10\beta} (10\beta)^x / x!$. It is a one-tailed test. Since a deletion may include multiple
 18 bases, the number of deleted bases, instead of the number of deletions, is used to
 19 determine the Poisson probability. If insertions are present, the window is ignored.
 20

21 In total, 330,942 high quality genomic sequences (submitted before 18 February, 2021)
 22 were analyzed and 27,042 windows were examined, and 12,930 windows containing
 23 significantly less mutations with an FDR-corrected P -value < 0.01 were found (Figure
 24 S5). Overlapped windows are merged to form a mutation cold spot (Supplemental
 25 excel file).

1 Estimation of mutation rate

2 Most SARS-CoV-2 strains were collected in different days. Similar to cases of
3 longitudinal samples (22), more mutations are accumulated subsequent to the
4 appearance of the most recent common ancestor (MRCA) (16).

5

6 Based on 178,765 high-quality genomic sequences submitted before 5 January, 2021,
7 linear regression was performed to estimate the mutation rate of SARS-CoV-2 (Figure
8 S6). For each strain with a different collection date in a tip-dated time tree, the
9 number of mutations, including that of recurrent mutations, was counted subsequent
10 to the appearance of MRCA. As described previously, recombination-flagged
11 mutations were excluded. Similar to the previous study (16), demography and the
12 time of MRCA appearance were not required for estimation of mutation rates.

13

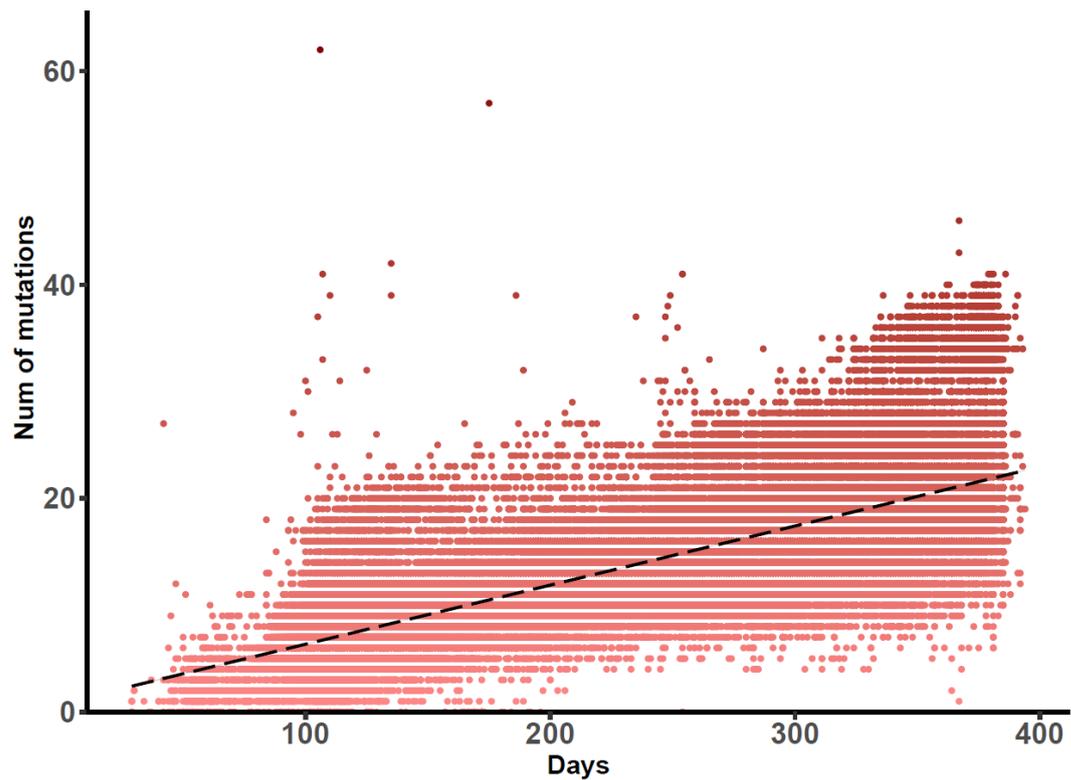
14 The regression line ($y = 0.0553x + 0.8086$) is obtained (Figure S6), where x is the
15 number of days between 1 December, 2019 and date of collection, and y is the
16 number of mutations accumulated since the appearance of MRCA of SARS-CoV-2.

17 The slop of the regression line indicates the genome-wide mutation rate of
18 SARS-CoV-2 (0.0553 per genome per day or 6.8017×10^{-4} per nucleotide per
19 year). As the extended regression line crosses the x -axis at -14.6 days, the time of
20 MRCA of SARS-CoV-2 was determined to be November 15, 2019. The mutation rate
21 of each SARS-CoV-2 gene is shown in Table S1.

22

23 The 95% confidence interval of the estimated mutation rate is obtained via
24 Monte-Carlo simulations. Given the estimated mutation rate, mutations are randomly
25 generated along the evolutionary tree (23), and mutation rate is estimated by
26 regression analysis. Then the empirical distribution of estimated mutation rate is
27 obtained from 1,000 simulated data set.

28



1

2 **Figure S6. Linear regression for estimation of mutation rate.** In this figure,
 3 178,765 high-quality genomic sequences (submitted before 5 January, 2021) were
 4 analyzed. The *x*-axis displays number of days between 1 December, 2019 and date of
 5 collection. The *y*-axis indicates number of mutations accumulated since the
 6 appearance of MRCA of SARS-CoV-2.

7

8

9 **Table S1. Mutation rate of various SARS-CoV-2 genes.**

10

SARS-CoV-2 Gene	Mutation rate (per nucleotide per year)
ORF1a	3.9707×10^{-4}
ORF1b	3.8675×10^{-4}
S	9.8983×10^{-4}
ORF3a	5.5584×10^{-4}
E	2.8996×10^{-4}
M	18.623×10^{-4}
ORF6	1.0830×10^{-4}
ORF7a	2.7202×10^{-4}
ORF7b	14.357×10^{-4}
ORF8	41.195×10^{-4}
N	18.458×10^{-4}
ORF10	75.992×10^{-4}
noncoding	29.448×10^{-4}

1 Maximum-likelihood phylodynamic analysis

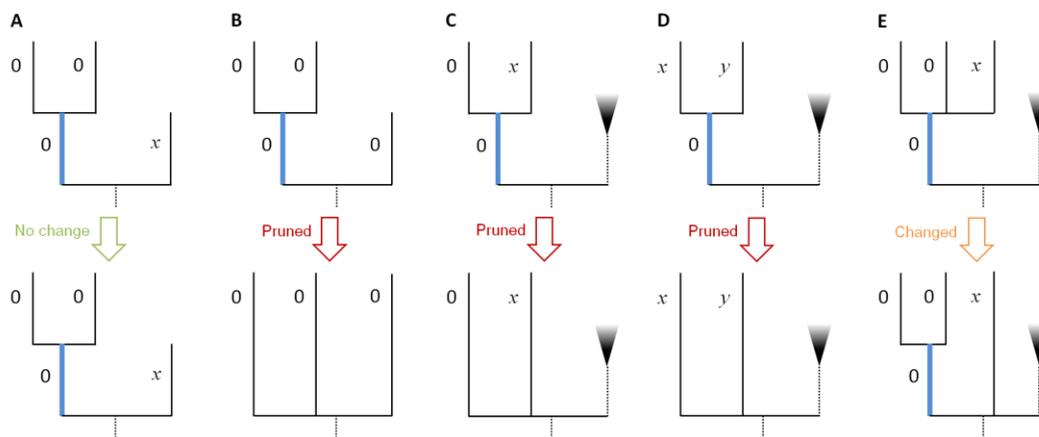
2 A highly effective maximum-likelihood method (TreeTime) is used to determine the
3 dates of internal nodes (24) as it allows fast inference by “the post- and pre-order
4 traversals” with tabulated key values for back tracing. This algorithm was
5 implemented in CGB with very minor revisions. The genome-wide mutation rate is
6 also timely updated to calculate the likelihood.

7

8 As recommended by TreeTime (24), all length zero branches are pruned, and branch
9 length corresponds to number of mutations on the branch. To improve computation
10 efficiency, CGB first categories branches with length zero according to its context
11 (Figure S7). In some cases, branches with length zero are not pruned (Figure S7A, E)
12 in order to make length zero offspring as a clade and to reduce the number of
13 multifurcated nodes.

14

15



16

17

18 **Figure S7. Five categories of length zero branches (highlighted in blue).**

19 A) All offspring of the branch have length zero, and the sister branch of the branch
20 has length non-zero x . In this case, the two offspring of length zero are in the same
21 clade.

22 B) The sister branch has length zero, and the three nodes are clustered to form a
23 multifurcated clade.

24 C) If one offspring of the branch has length zero, the branch is pruned.

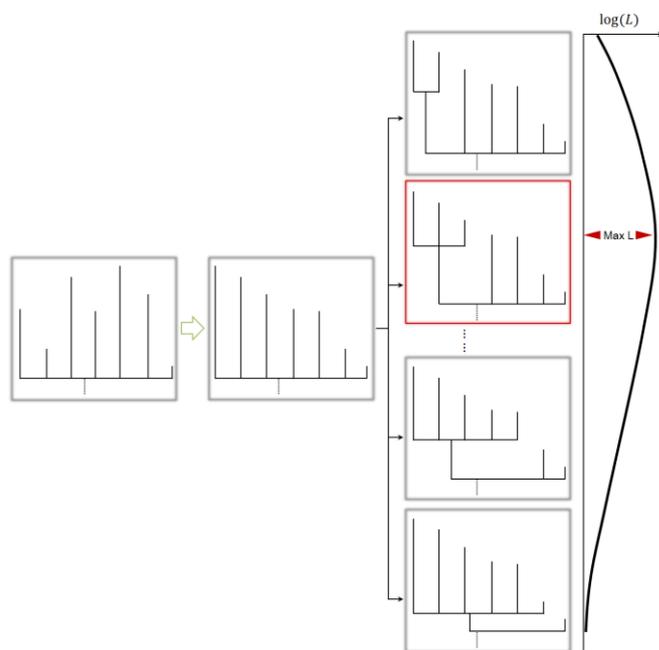
25 D) If all offspring of the branch have length non-zero x or y , the branch is pruned.

26 E) If two or more offspring of the branch have length zero, the branch is kept and the
27 non-zero branch is removed.

28

29

1 Many internal nodes are multi-furcated instead of bi-furcated because the viral strains
 2 are very similar to each other. The multi-furcated nodes are known as polytomies. To
 3 reduce the number of branches of a polytomous node, CGB sorts the branches
 4 according to the potential gain of likelihood if branches are shortened and determines
 5 whether a longer or shorter branch length would increase the likelihood of tree. The
 6 branches are bi-partitioned to form a new clade (Figure S8), and the two sets of
 7 branches are determined by maximizing the gain of likelihood. The bi-partition
 8 always starts from the root to the tips, and this process is repeated at least four times.
 9 The date of the MRCA appearance was estimated by this method to be 15 November,
 10 2019 (95% CI: 21 October – 22 November, 2019), the same as that estimated by the
 11 regression analysis described above. As the first onset of COVID-19 was reported on
 12 8 December, 2019 (25), the result suggests that viral spread and evolution occurred
 13 before that date.
 14



15
 16 **Figure S8. Bi-partition of a polytomous node.** CGB first sorts the branches according
 17 to the potential gain of likelihood. If k branches are linked to the node, there are
 18 $k - 2$ different ways to bi-partition the node. The two sets of branches are
 19 determined by maximizing the gain of likelihood.

20 **Displaying SARS-CoV-2 genomic variants in tree-based file format**

21 Similar to NextStrain (26) and the WashU Virus Genome Browser (27), CGB uses a
 22 tree-based file format to show SARS-CoV-2 genomic variants. The head of data file
 23 contains data version, update date, genomic region analyzed, and mutation rate
 24 estimated for each gene. The data file is in Newick tree format (nwk) and contains
 25 information on collection date, gender and age of patient, location for each strain,

1 mutations, and inferred internal nodes. Recombination flags are not included in the
2 output data because they can be easily reconstructed. To allow fast showing and
3 re-analyzing large number of SARS-CoV-2 genomic variants, redundant data are
4 minimized.

5 **Maximum-likelihood analysis based on the existing mutation-annotated tree**

6 Branch and bound for maximum parsimony (28, 29) is implemented with a speed-up
7 revision. New genomic sequences of SARS-CoV-2 strains are first aligned with the
8 reference genome (Figure 1). The resulting alignment and previous results are then
9 analyzed together, and the evolutionary tree is rebuilt using previous result file that
10 contains the existing tree and mutation information. A new strain is then added to the
11 mutation-annotated tree as a dated leaf, and new mutations are labeled and analyzed
12 according to the principle of parsimony. CGB adds the earliest strain first to the tree.
13 After adding all new genomic sequences, the mutation rate of SARS-CoV-2 is
14 calculated, and the date of each internal node is determined as described above. This
15 maximum-likelihood (ML) analysis was performed with a slightly revised version of
16 TreeTime (24).

17
18 The speed-up-revised branch and bound provides a balance between efficiency and
19 accuracy. However, it may not be globally optimized. To solve this problem, a
20 sub-tree optimization is performed. As many internal branches have five or more
21 mutations, the large evolutionary tree was divided into small subtrees. Because
22 sub-tree optimization is much faster than rebuilding the whole tree, it is frequently
23 performed as needed.

24 **CGB binary nomenclature for each internal node or branch**

25 A number of different naming systems have been proposed (26, 30, 31), but these
26 systems only name a very small number of internal nodes or branches. As there are
27 98,496 internal nodes on the huge evolutionary tree of 330,942 SARS-CoV-2 strains,
28 it is nearly impossible to manually label each node. Therefore, the CGB binary
29 nomenclature system was developed following the MRCA concept as follows.

30

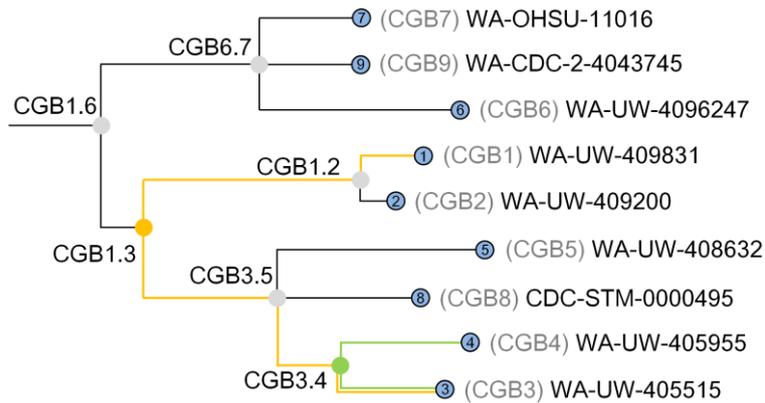
31 Each node of a viral strain is first assigned a permanent unique positive integer (e.g., 1
32 – 9) in the order of discovery (Figure S9). Assuming that an internal node has two
33 sub-nodes that are named CGB1 and CGB2, this internal node is named CGB1.2. For
34 an internal node with more than two sub-nodes, e.g., CGB7, CGB9, and CGB6, it is
35 named with the two smallest CGB numbers, given the condition that the internal node

1 is the MRCA of the two sub-nodes, separated by a dot; thus, this internal node is
2 designated as CGB6.7.

3

4 This naming process is very fast, and all nodes of the huge evolutionary tree can be
5 named in seconds. Each node can be easily searched and viewed by CGB. When a
6 new sequence is added to the tree as a sub-node, its CGB number would be greater
7 than all the pre-existing CGB numbers and thus will not change the previously
8 assigned CGB number of the internal node, which the new sequence belongs to.

9



10

11

12 **Figure S9. Illustration of CGB binary nomenclature.** The evolutionary tree is
13 shown with nine strains indexed CGB1 – CGB9. Their pseudo-isolate names are
14 shown. The green internal node with two sub-nodes named CGB3 and CGB4 is
15 designated as CGB3.4 since the MRCA of CGB3 and CGB4 is the green node. For an
16 internal node with more than two sub-nodes, it is named with the two smallest CGB
17 numbers, given the condition that the internal node is the MRCA of the two sub-nodes,
18 separated by a dot. Therefore, an orange internal node is named CGB1.3 because it
19 contains CGB1, CGB2, CGB5, CGB8, CGB4, and CGB3 with CGB1 and CGB3
20 being the smallest CGB numbers, on the condition just described.

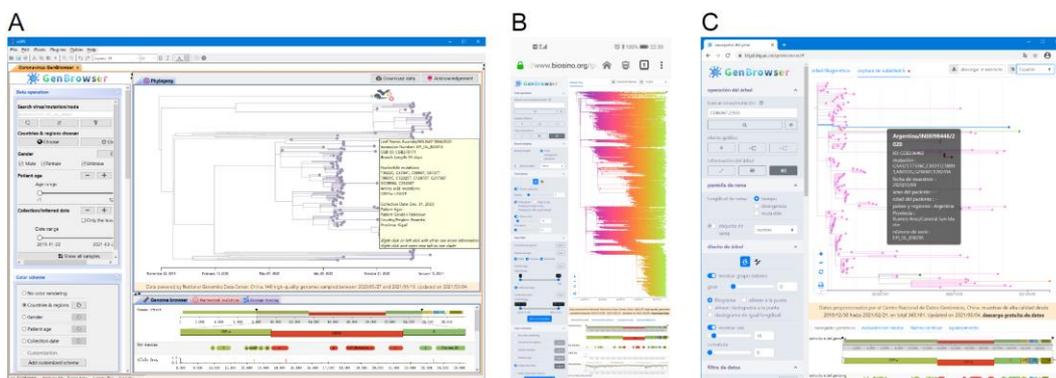
21 Visualization of the huge evolutionary tree by movie-maker strategy

22 When visualizing the huge evolutionary tree, many lineages are invisible because they
23 overlap each other. If the height of a drawing panel is 1,000 pixels and 10,000
24 horizontal lines are needed to visualize the entire set of data, only 1,000 surface lines
25 can be seen. The other 9,000 lines are invisible because they are located below the
26 surface lines. Therefore, it is not necessary to paint the 9,000 invisible lines. If the tree
27 is zoomed in to show details, only a sub-area of the tree is painted and visible. With
28 this strategy, hundreds of thousands of lineages can be visualized effectively even on
29 a smart phone (Figure S10B).

1 **Data searching, filtering, and visualization of a single clade on the huge tree**

2 To view a lineage on the huge evolutionary tree, several different data searching and
3 filtering methods can be used. A clade can be viewed in a new tab, and its sub-clade
4 can be viewed in another new tab. A clade can also be collapsed or un-collapsed.
5 Moreover, chosen lineages can be made visible, and un-chosen ones can be hidden.
6 After right clicking a branch, a menu will pop up to help navigate through the huge
7 tree. A lineage can also be viewed by deep zoom-in using the desktop standalone
8 version of CGB. However, the deep zoom-in function is not implemented in the
9 web-based CGB because it is a simplified version and is designed mainly for
10 educational purpose.

11 **Tree visualization with CGB**



12
13 **Figure S10. Tree visualization with CGB.**

14 A) Tree visualization of 148 SARS-CoV-2 strains collected from Rwanda, Africa.
15 B) Web-based CGB tree visualization of 360,181 genomes with the Android version
16 of Firefox.
17 C) Web-based CGB tree visualization of an Argentinian clade (CGB6867.22533) in
18 Spanish with the desktop version of Google Chrome. Nine language versions
19 (Chinese, English, German, Japanese, French, Italian, Portuguese, Russian, and
20 Spanish) are available.

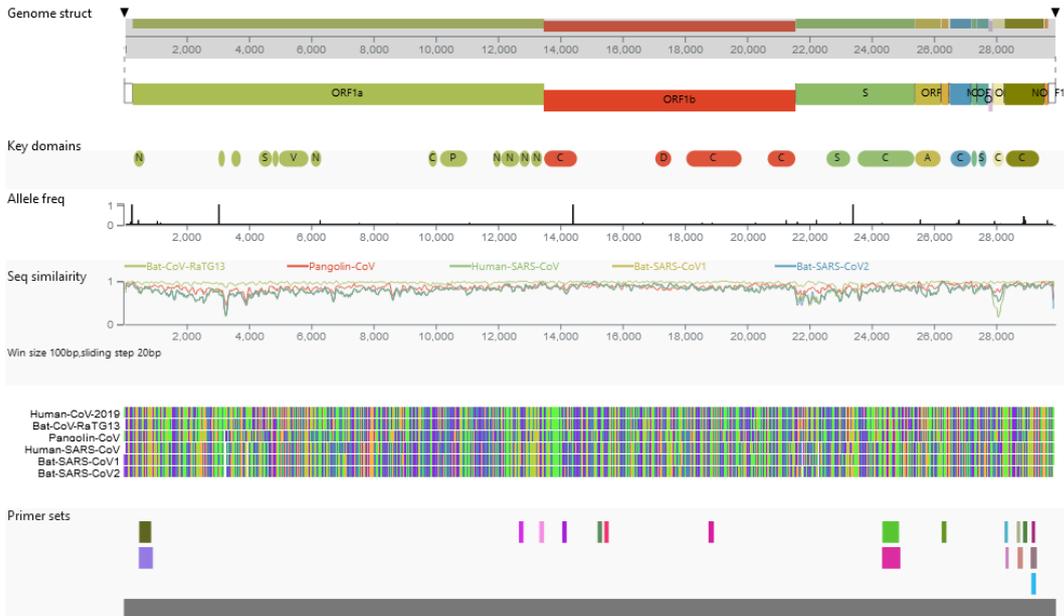
21 **Coordinated annotation tracks**

22 CGB uses six tracks to show genome structure and key domains; allele frequencies;
23 sequence similarity between various coronavirus isolated from human, bat, and
24 pangolin; multi-genome alignment; and primer sets for detection of various

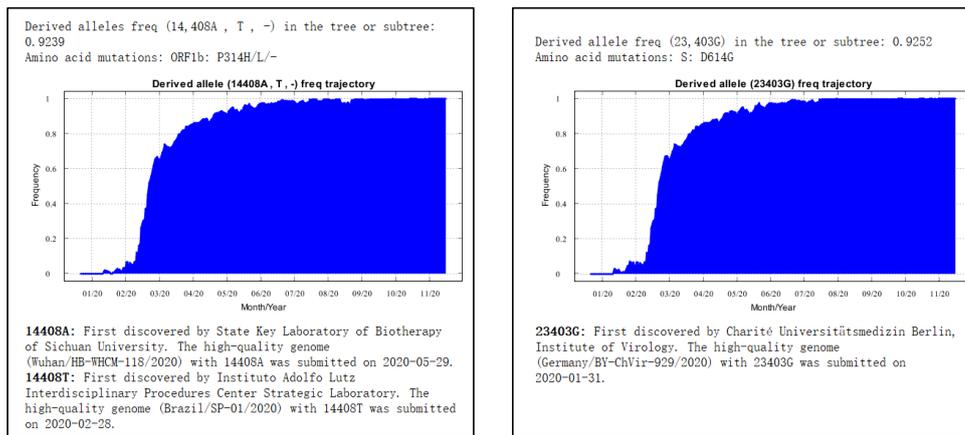
1 SARS-CoV-2 genes and strains (Figure S11). These tracks are coordinated according
2 to nucleotide positions of the SARS-CoV-2 reference genome.

3
4 The first track shows the structure of a SARS-CoV-2 genome. By dragging or right
5 clicking the mouse, a genomic region can be zoomed in. The second track shows 25
6 known key domains. By right clicking on a domain box, amino acid sequence of the
7 domain can be copied, and the related information page on the Pfam website
8 (<http://pfam.xfam.org>) can be opened.

9
10 The third track shows the frequencies of derived alleles or variants (Figure S11).
11 Since the web version is designed for the general public and quick view of global
12 samples, users can update manually the frequency of an allele in the chosen clade.
13 When hovering mouse on the frequency column of an allele, its allele frequency
14 trajectory (Figure S12) will pop up. This allele frequency trajectory is calculated by a
15 sliding window of five days in size. The person who first discovered the allele is
16 indicated below allele frequency trajectory.



17
18 **Figure S11. Six tracks shown by the Coronavirus GenBrowser.**
19



1

2 **Figure S12. Visualization of allele frequency trajectory with CGB.**

3

4 The fourth track shows (14408A, T, -) sequence similarity between SARS-CoV-2 reference genome
5 (NC_045512.2) (3) and the genomes of five other coronaviruses, including
6 bat-CoV-RaTG13 (MN996532.1) (4), pangolin-CoV (MT040334.1) (5),
7 human-SARS-CoV (AY278488.2) (32), bat-SARS-CoV1 (KY417146.1) (33), and
8 bat-SARS-CoV2 (MK211376.1) (34). Sequence similarity is determined using a
9 sliding window (window size 100 bp and sliding step 20 bp). In the standalone
10 version of CGB, these parameters can be adjusted to re-calculate the degree of
11 sequence similarity.

12

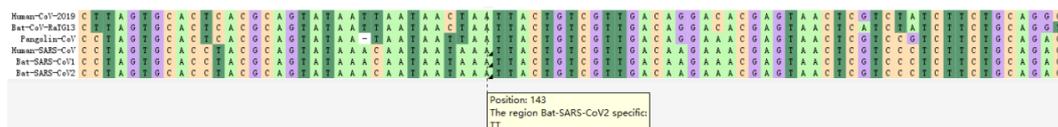
13 The fifth track shows alignments of six coronaviruses performed with the software
14 MAFFT (2). Nucleotide sequences of five coronaviruses are coordinated according to
15 nucleotide positions of the SARS-CoV-2 reference genome. Inserted sequences, if any,
16 in the genomes of the five non-SARS-CoV-2 coronaviruses can be viewed with the
17 standalone version of CGB (Figure S13).

18

19 The sixth track presents primer sets that can be used to detect various SARS-CoV-2
20 genes or strains. Various regions of the genome that can be amplified are indicated.
21 Combined with allele frequency information, the efficiency of nucleic acid testing can
22 be verified (Figure S14). Since viral strains can be filtered according to collection
23 dates and locations, their allele frequency can be easily determined.

24

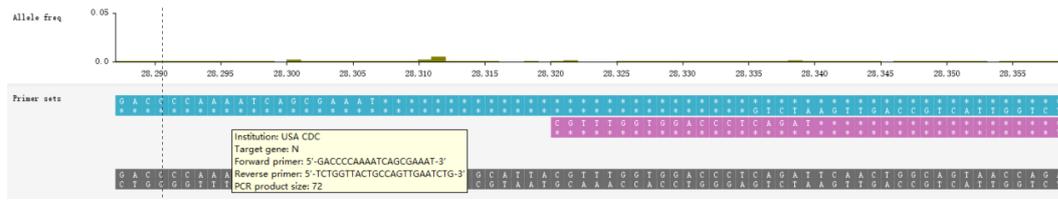
25



26

27 **Figure S13. Multiple-genome alignment.** Inserted sequences in the five
28 non-SARS-CoV-2 genomes are marked with black triangles. This alignment can be
29 downloaded from <https://bigd.big.ac.cn/ncov/apis/>.

1



2

3

4 **Figure S14. Combined view of two tracks of allele frequency and primer set for**
 5 **detection of various SARS-CoV-2 strains.** The nucleotide sequences of two primers
 6 are shown, and their amplified region is marked in pink.

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

8 To detect branch-specific accelerated evolution, each internal branch of the
 9 SARS-CoV-2 tree was examined. For each internal branch, the observed number of
 10 mutations of the i -th gene ($\gamma_{obs,i}$) was compared with the expected number of
 11 mutations of the same gene ($\gamma_{exp,i}$). The significance level of acceleration was
 12 determined by Poisson probability (20, 21):

13 $P(x \geq \gamma_{obs,i}) = \sum_{x \geq \gamma_{obs,i}} e^{-\gamma_{exp,i}} (\gamma_{exp,i})^x / x!$, where $\gamma_{exp,i} = t\mu_i l_i / 365$, t is the
 14 duration (in days) of the branch, μ_i is the mutation rate (per nucleotide per year) of
 15 the i -th gene calculated from the entire collection of viral strains (Table S1), and l_i is
 16 the length of the i -th gene. It is a one-tailed test. The condition $t > 10$ (days) was
 17 used for detection of branch-specific accelerated evolution of SARS-CoV-2.

18

19 There were 61,561 internal branches with $t > 10$ (days) on the evolutionary tree
 20 ($n = 330,942$), and genome-wide accelerated evolution was detected on 36 branches
 21 (FDR corrected $P < 0.05$). Accelerated evolution of ORF1ab was found on 210
 22 branches (FDR corrected $P < 0.05$), and that of the spike gene was not observed (FDR
 23 corrected $P < 0.05$) (Supplemental excel file).

24

25 The internal branch CGB267525.267526 (genome-wide FDR corrected $P =$
 26 1.2302×10^{-10}) that had the most accelerated evolution was dated November 2020
 27 with only two descendants. To determine whether variants with accelerated evolution
 28 were advantageous in spreading, the median number of descendants of other variants
 29 in the same time period (± 10 days) were determined and found to be seven. The
 30 probability for the CGB267525.267526 variant to have two or more than two
 31 descendants is 0.7762 (one-tailed). Thus the accelerated variant may not be more
 32 contagious than others.

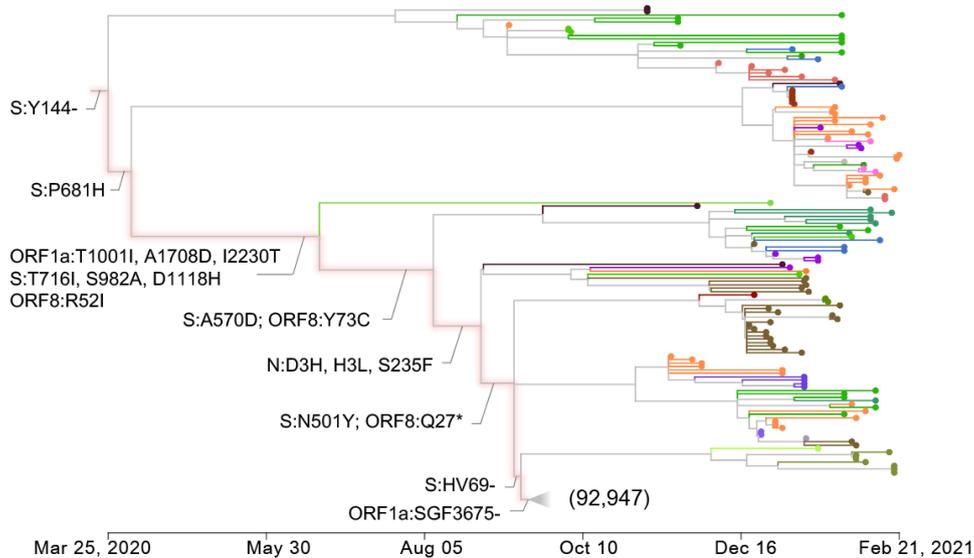
33

1 In total, 225 variants with accelerated evolution were detected. However, all the
2 evolution-accelerated variants were not found to spread significantly faster than other
3 variants during the same period of time. The number of descendants of the majority
4 (168/225=74.6%) of these variants was lower or equal to the median number of
5 descendants of variants with no accelerated evolution, suggesting that these
6 accelerated evolutions are mostly neutral.

7 Sequential occurrence of B.1.1.7-associated mutations

8 The variant of concern B.1.1.7 (CGB66679.75056) identified in the UK was recently
9 reported (35). The CGB evolutionary tree shows the sequential occurrence of
10 B.1.1.7-associated non-synonymous mutations (Figure S15). The results indicate that
11 S:HV69-, and ORF1a:SGF3675- were recently occurred.

12
13



14

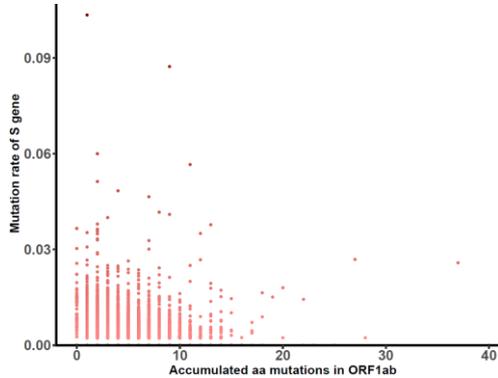
15 **Figure S15. Tree visualization showing the sequential occurrence of**
16 **B.1.1.7-associated non-synonymous mutations.**

17 No correlation between ORF1ab and spike gene mutation rates

18 ORF1ab encodes a polyprotein that is involved in genome transcription and
19 replication. Analysis was performed to determine whether the number of accumulated
20 non-synonymous mutations in ORF1ab correlates with the mutation rate of spike gene.
21 For each strain, the number of non-synonymous mutations in ORF1ab accumulated
22 after the appearance of MRCA of SARS-CoV-2 was determined. The spike mutation

1 rate was estimated based on the number of nucleotide mutations accumulated after the
2 appearance of MRCA of SARS-CoV-2 divided by the length of duration (in days). No
3 correlation between ORF1ab and spike gene mutation rates was observed (Pearson's
4 correlation coefficient = -0.00929) (Figure S16).

5



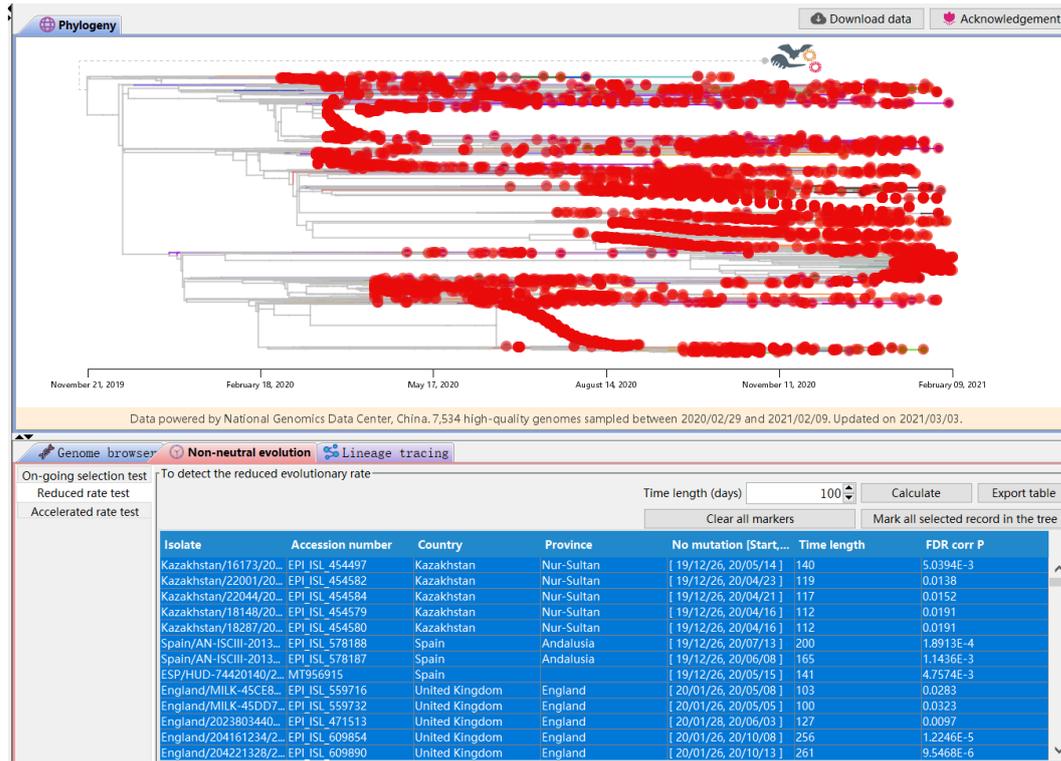
6

7 **Figure S16. Correlation between number of ORF1ab non-synonymous mutations**
8 **and spike gene mutation rate.** The number of high quality SARS-CoV-2 genomic
9 sequences is 330,942.

10 **Detection of SARS-CoV-2 strains that evolve more slowly recently**

11 CGB can also determine whether a SARS-CoV-2 strain evolves more slowly recently.
12 Based on 178,765 high-quality SARS-CoV-2 genomic sequences, the genome-wide
13 mutation rate was determined to be $\mu = 6.8017 \times 10^{-4}$ per nucleotide per year. It
14 indicates that a mutation occurs in a strain every 18.08 days. To detect variants with
15 reduced evolution rate, CGB determines the number of days during which a chosen
16 strain did not mutate. With Poisson probability (21), CGB also determines whether
17 the number of mutations of the strain is fewer than that expected during the time
18 period (one-tailed) (Figure S17). Detailed results of this analysis on three closely
19 related internal branches are shown in Figure 4A.

20



1
2

3 **Figure S17. SARS-CoV-2 strains with reduced evolution rate among 330,942**
4 **SARS-CoV-2 genomic sequences.** Each red dot represents a strain that evolved more
5 slowly recently. The table shows some examples of such strains. The three analyzed
6 closed related branches (Figure 4A) are marked with a blue empty circle in the upper
7 panel.

8 **Detection of on-going selection of SARS-CoV-2**

9 To detect on-going positive selection, allele frequency trajectory with an S-shaped
10 curve is examined as described previously (36-39). For this determination, the
11 selection coefficient is denoted as s . The initial frequency of the derived allele a is
12 denoted as q_0 , and that of the wide-type allele A is $p_0 = 1 - q_0$. The frequency of
13 the wild-type allele A at a specific day (time t) is p_t , and that of the derived allele a is
14 q_t .

15

16 The following equation was used to calculate the coefficient of on-going positive
17 selection (36) (Table S2):

18
$$\frac{q_t}{p_t} = (1 + s) \frac{q_{t-1}}{p_{t-1}} = \dots = (1 + s)^t \frac{q_0}{p_0} \quad (1)$$

19 Then

20
$$\log \left(\frac{q_t}{p_t} \right) = \log \left(\frac{q_0}{p_0} \right) + t \log(1 + s). \quad (2)$$

1 Since t is known, $\log(1 + s)$ can be estimated by linear regression.

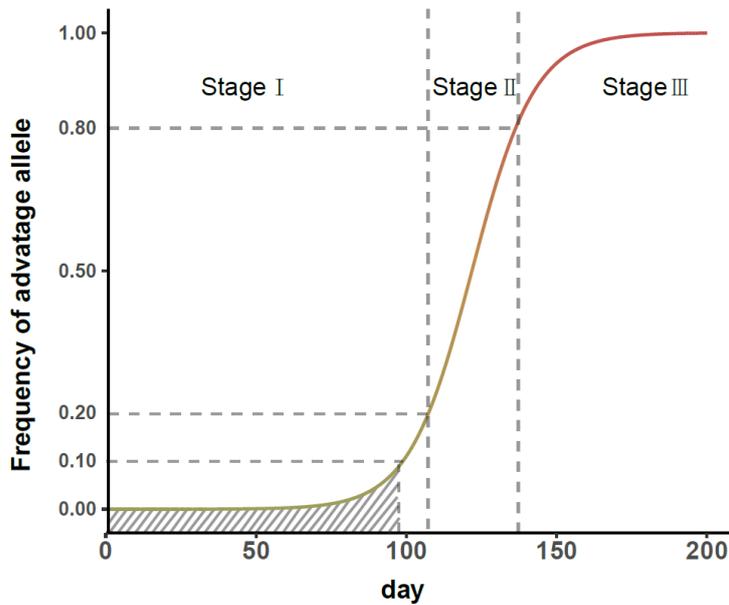
2

3 **Table S2. Frequency of wild type and derived alleles after selection.**

Haplotype	A (wild type)	a (derived allele)
Fitness	1	$1 + s$
Frequency at the $(t - 1)$ -th day	p_{t-1}	q_{t-1}
Frequency at the t -th day	$p_t = \frac{p_i}{p_i + (1 + s)q_i}$	$q_t = \frac{(1 + s)q_i}{p_i + (1 + s)q_i}$

4

5



6

7 **Figure S18. S-shaped frequency trajectory of advantageous mutations.** $s = 0.1$
 8 and $q_0 = 0.0001$. The best time window to control the transmission of a strain with
 9 an advantageous mutation is shadowed.

10

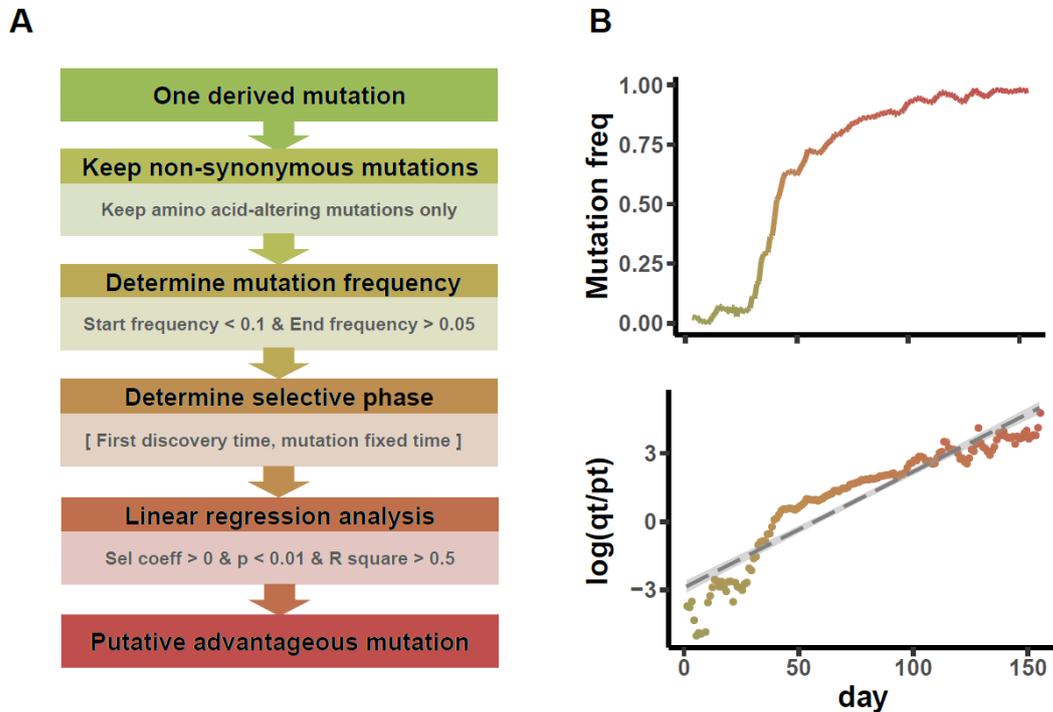
11 When $s > 0$, the frequency of a derived allele increases over time (36) (Figure S18).
 12 During Stages I and III, the speed of increase in the frequency of advantageous allele
 13 is slow, indicating low selection efficiency. During Stage II, the speed of increase in
 14 the frequency of advantageous allele is fast, and the efficiency of selection is high.
 15 When the frequency is 50%, the efficiency of selection reaches maximum. Therefore,
 16 the best time window to control the transmission of strains with an advantageous
 17 mutation is that of Stage I, especially when its frequency is still below 10%.

18

19 The analysis framework for detecting strains with putative advantageous mutations
 20 during their early stage of spreading is summarized in Figure S19. A neutral mutation
 21 may be linked to an advantageous mutation and spread over the entire population (14,
 22 15, 40). To reduce the impact of hitchhiking by neutral mutation, only
 23 non-synonymous mutations were analyzed. For this analysis, the initial (start)
 24 frequency must be < 0.1 , and the end frequency must be > 0.05 . Only the mutation

1 frequency trajectory during the selective phase was used for calculation as this is the
 2 period when an advantageous mutation causes on-going selection. Linear regression
 3 analysis was performed to detect advantageous mutation. According to the equation
 4 described above, a mutation was considered advantageous when $s > 0$, $p < 0.01$,
 5 and $R^2 > 0.5$.

6



7

8 **Figure S19. Detection of on-going selection of SARS-CoV-2.**

9 **A)** Flow chart for detection of putative advantageous variants.

10 **B)** Frequency trajectory for A23403G (S: D614G) and linear regression analysis. The
 11 x -axis displays number of days since the first appearance of a derived allele in global
 12 virus population. q_t is the frequency of the derived allele (23403G), and p_t is the
 13 frequency of the ancestral allele (23403A) at time t .

Table S3. Putative advantageous mutations in the spike protein.¶

Position	Nucl. mut.	AA mut.	Start time	Start freq	End/Last time	End/Last freq	Sel Coeff	P-value	R-square
21765*	TACATG21765-	HV69-	2020/3/26	0.0003	2021/2/8	0.7716	0.0321	<1.0E-10	0.9374
21991*	TTA21991-	Y144-	2020/3/1	0.0031	2021/2/8	0.7776	0.0223	<1.0E-10	0.6073
23063*	A23063T	N501Y	2020/3/28	0.0002	2021/2/8	0.7731	0.0346	<1.0E-10	0.7856
23271*	C23271A	A570D	2020/4/25	0.0004	2021/2/8	0.7687	0.0566	<1.0E-10	0.7843
23403	A23403G	D614G	2020/1/22	0.0208	2020/7/21	0.9917	0.0437	<1.0E-10	0.8615
23604*	C23604A	P681H	2020/1/28	0.0056	2021/2/8	0.794	0.0279	<1.0E-10	0.8
23709*	C23709T	T716I	2020/3/28	0.0013	2021/2/8	0.7836	0.0295	<1.0E-10	0.7143
24506*	T24506G	S982A	2020/9/18	0.0005	2021/2/8	0.7716	0.077	<1.0E-10	0.9326
24914*	G24914C	D1118H	2020/3/31	0.0002	2021/2/8	0.7642	0.0537	<1.0E-10	0.7633

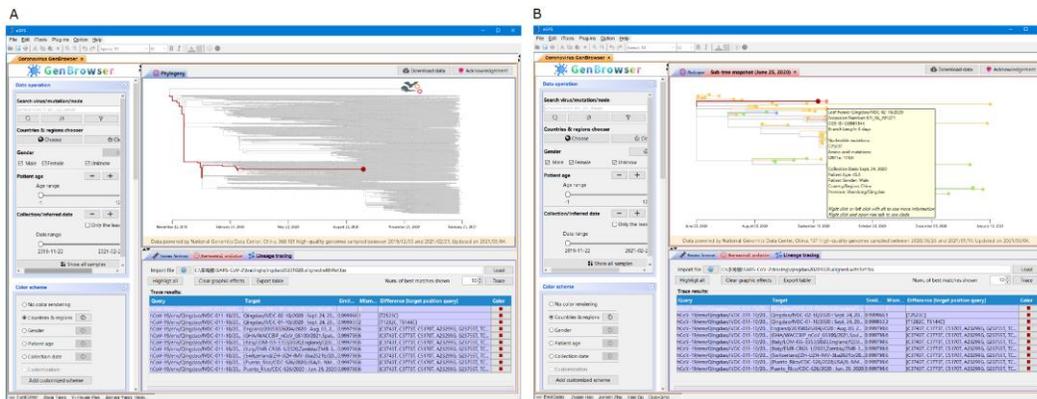
¶The analysis was performed on global samples ($n = 330,942$) submitted up to Feb. 18, 2021. Functions of D614G have been characterized. (41, 42)

*Mutations found on clade CGB66679.75056 (B.1.1.7).

1 **Lineage tracing**

2 For lineage tracing (Figure S20), genomic sequences of SARS-CoV-2 strains (query)
3 collected from patients or environments are aligned with the genomic sequence of the
4 most recent common ancestor (MRCA) of SARS-CoV-2 (GenBank accession number:
5 NC_045512) (3). Lineage-specific mutations are then inferred for each query strain and
6 the strain of each node on the evolutionary tree. Two sets of mutations are compared, and
7 the node with the least difference from the query is considered as the target node, which
8 may be a leaf or an internal node.

9
10 With CGB, lineage tracing of the entire collection of 360,181 strains can be completed in
11 seconds if aligned sequences are used as input. A similar approach (Ultrafast sample
12 placement on existing trees, UShER) was independently developed recently (43) and
13 published on bioRxiv on 28 September, 2020, while CGB lineage tracing was first
14 released online (<http://www.egps-software.net/>) on 12 August, 2020.



17
18 **Figure S20. Lineage tracing with CGB.**

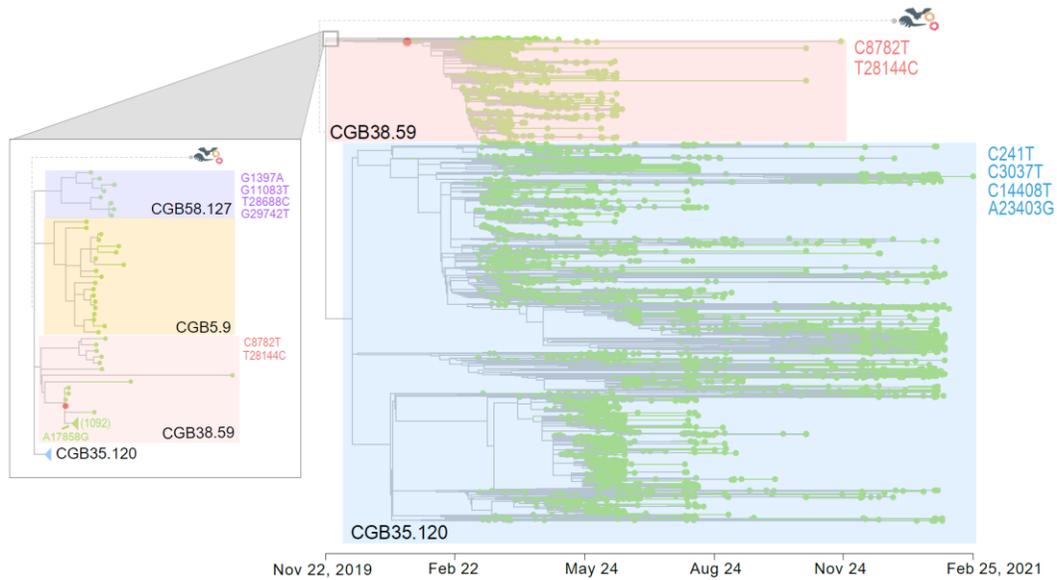
19 A) Global view of the traced lineage for Qingdao outbreak in a huge SARS-CoV-2
20 evolutionary tree ($n = 330,942$).

21 B) Zoom-in view of the traced lineage for Qingdao outbreak in the SARS-CoV-2
22 evolutionary tree.

1 Analysis of SARS-CoV-2 transmission in Washington State

2 The first outbreak of COVID-19 in Washington State was analyzed using the sequences
3 of 453 SARS-CoV-2 isolates collected between 20 February and 15 March, 2020 (6).
4 Results suggest that this outbreak was derived from a single introduction. Since then,
5 more isolates were collected, and a total of 5,170 high-quality genomic sequences were
6 obtained. Results of phylogenetic studies showed that these 5,170 strains belong to four
7 clades (Figure S21). Therefore, deep sequencing could provide more details in the
8 analysis of genomic epidemiology (44).

9
10 The CGB38.59 clade described before (6) contains 1,110 strains with the following
11 mutations: C8782T and T28144C. The virus isolated from the first COVID19 case in the
12 United States (45) is marked in red in Figure S21. This strain has single-nucleotide
13 polymorphisms (SNPs) C8782T, C18060T, and T28144C compared to the MRCA of
14 SARS-CoV-2. The CGB35.120 clade contains 4,033 strains. All strains in this clade have
15 the following mutations: C241T, C3037T, C14408T, and A23403G. Strains in this clade
16 are prevalent ($4,033 / 5,170 = 78.0\%$) probably because they carry the advantageous
17 mutation D614G (A23403G) in the spike protein (41). Two minor clades (CGB5.9 and
18 CGB58.127) are linked to the root of the tree. These clades are shown in the sub-panel
19 (Figure S21) when major clades are collapsed.
20



1

2 **Figure S21. Phylogeny of 5,170 high-quality SARS-CoV-2 genomic sequences of**
 3 **strains found in Washington State.**

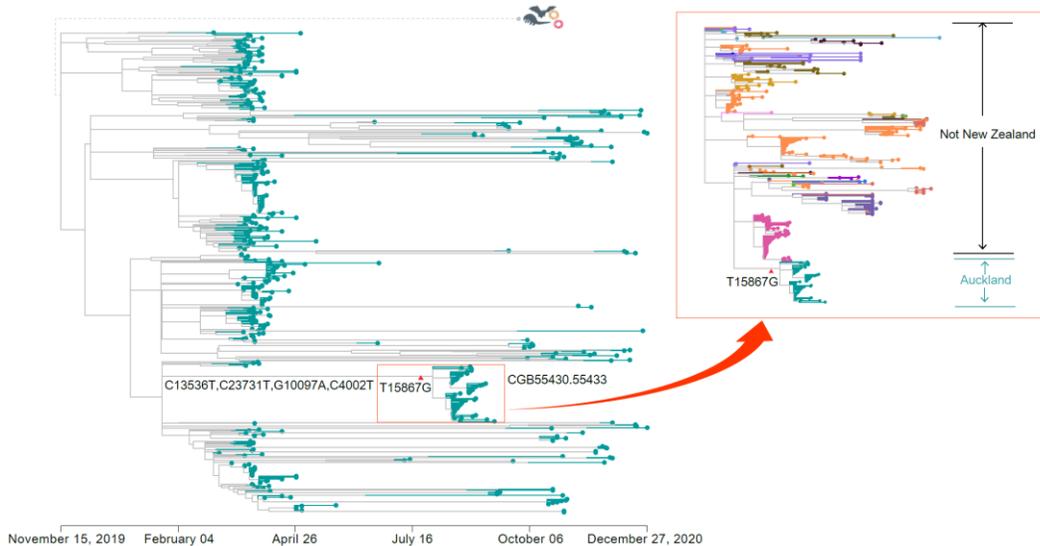
4 The virus isolated from the first COVID19 case in the United States (45) is marked with a
 5 red circle (in CGB38.59 clade). This strain has single-nucleotide polymorphisms (SNPs)
 6 C8782T, C18060T, and T28144C compared to the MRCA of SARS-CoV-2. The
 7 mutations that distinguish these clades are indicated. The sub-panel shows enlarged view
 8 of the two minor clades.

9 **Phylogenetic analysis of strains in the Auckland outbreak**

10 There was an outbreak in Auckland in August 2020 after more than 100 days without a
 11 local transmission of COVID-19 in New Zealand. The first person who was tested
 12 positive in this outbreak worked at Americold food cold-storage facility in Auckland.
 13 CGB phylogenetic analysis revealed that all 56 strains in the CGB555430.55433 clade
 14 isolated from this outbreak carry a novel mutation T15867G, suggesting that the outbreak
 15 was derived from a single strain. This variant, as a genomic signature, was only found in
 16 strains associated with the Auckland outbreak (Figure S22).

17

18 Four mutations (C13536T, C23731T, G10097A, and C4002T) present in strains of the
 19 Auckland outbreak are not found in any other strains collected from other outbreaks in
 20 New Zealand (Figure S22). However, these four mutations are found in many strains
 21 collected from other countries. These observations suggest that the strains responsible for
 22 the outbreak in Auckland are different from those of the first outbreak in New Zealand in
 23 April 2020.



1
2

3 **Figure S22. Phylogeny of 483 high-quality SARS-CoV-2 genomic sequences of**
 4 **strains found in New Zealand.** The CGB55430.55433 clade of strains from the
 5 Auckland outbreak is highlighted. The mutations that distinguish strains of the Auckland
 6 outbreak from those of other outbreaks in New Zealand are indicated. The subpanel
 7 shows enlarged view of strains found in the Auckland outbreak together with their
 8 relatives found in other countries harboring the same mutations (C13536T, C23731T,
 9 G10097A, and C4002T).

10 **Acknowledgement of the person who first discovers any SARS-CoV-2 variant**

11 To encourage data sharing, CGB acknowledges the person who first discovered a specific
 12 strain (Figure S7), and such information can be easily found with CGB.

13 **D614G mutation in the spike protein**

14 Results of CGB analysis suggest that the D614G mutation occurred between 10
 15 November and 9 December, 2019. This mutation was first discovered by Institute of
 16 Virology, Charité Universitätsmedizin Berlin, and the sequence of the first strain
 17 (Germany/BY-ChVir-929/2020) with this mutation was deposited publicly on 31 January,
 18 2020 (Figure S12).

1 Standalone and web-based CGB

2 The web-based CGB is a simplified version in multiple languages and can be accessed
3 with any web browser to view the tree, search a viral strain or a mutation, and perform
4 data filtering. However, it does not monitor non-neutral evolutions (including accelerated
5 and reduced evolution) and perform lineage tracing. It also does not have sufficient speed
6 for viewing allele frequency trajectory and cannot zoom in to view an individual lineage.
7 These functions are available in the standalone alone CGB. The standalone alone CGB is
8 a plug-in module for the eGPS software (<http://www.egps-software.net/>) (46). It provides
9 the full function of CGB, and allele frequency trajectory can be promptly obtained.

10 CGB in multiple languages

11 The web-based CGB is written in eight different languages, including Chinese, English,
12 German, French, Italian, Portuguese, Russian, and Spanish. Therefore, the general public
13 in many regions of the world can easily access timely pre-analyzed results of the latest
14 SARS-CoV-2 genomes. Different language versions are implemented with different
15 configuration files (Table S4). These configuration files are available upon request and
16 can be freely translated into other languages.

17

18

19 **Table S4. Examples of configuration files in different languages.**

Variable name	English	Chinese	German
app_name	genbrowser	新冠病毒基因组浏览器	GenBrowser
download_desktop	Download Desktop	下载桌面版程序	Desktop herunterladen
tree_ope	Tree operation	进化树操作区域	Operationsbereich des Baums
search_node	Search virus/mutation	查找病毒/变异	Virus/Mutation suchen
institution	institution	组织机构	Institution
forward_primer	Forward primer	正向引物	Vorwärtsprimer
reverse_primer	Reverse primer	反向引物	Rückwärtsprimer

1 Timely update of CGB

2 CGB provides timely or daily updates as needed.

3 Statistical information

4 The tests for detecting mutation cold spots, branch-specific accelerated evolution of
5 SARS-CoV-2 are one-tailed. One-tailed test was applied for detecting whether a variant
6 is advantageous in spreading.

7 Data availability

8 The coronavirus genomic sequences used in this study were obtained from the
9 2019nCoV database (1). Timely updated data of genomic sequences of SARS-CoV-2
10 variants are shared with the general public at <https://bigd.big.ac.cn/ncov/apis/>. The free
11 software (desktop and web-based versions) can be downloaded from
12 <http://www.egps-software.net/>.

13 References

- 14 1. Zhao, W-M, Song, S-H, Chen, M-L, *et al.* The 2019 novel coronavirus resource. *Hereditas (Beijing)*.
15 2020; **42**(2): 212-21.
- 16 2. Rozewicki, J, Li, S, Amada, KM, *et al.* MAFFT-DASH: integrated protein sequence and structural
17 alignment. *Nucleic Acids Res.* 2019; **47**(W1): W5-W10.
- 18 3. Wu, F, Zhao, S, Yu, B, *et al.* A new coronavirus associated with human respiratory disease in China.
19 *Nature.* 2020; **579**: 265-9.
- 20 4. Zhou, P, Yang, X-L, Wang, X-G, *et al.* A pneumonia outbreak associated with a new coronavirus of
21 probable bat origin. *Nature.* 2020; **579**: 270-3.
- 22 5. Lam, TT-Y, Jia, N, Zhang, Y-W, *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan
23 pangolins. *Nature.* 2020; **583**(7815): 282-5.
- 24 6. Bedford, T, Greninger, AL, Roychoudhury, P, *et al.* Cryptic transmission of SARS-CoV-2 in
25 Washington state. *Science.* 2020; **370**(6516): 571-5.
- 26 7. Price, MN, Dehal, PS, Arkin, AP. FastTree 2-Approximately maximum-likelihood trees for large
27 alignments. *PLoS ONE.* 2010; **5**(3): e9490.

- 1 8. Kozlov, AM, Darriba, D, Flouri, T, *et al.* RAxML-NG: a fast, scalable and user-friendly tool for
2 maximum likelihood phylogenetic inference. *Bioinformatics*. 2019; **35**(21): 4453-5.
- 3 9. Saitou, N, Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees.
4 *Mol Biol Evol*. 1987; **4**(4): 406-25.
- 5 10. Li, H, Zhang, YW, Zhang, YP, *et al.* Neutrality tests using DNA polymorphism from multiple samples.
6 *Genetics*. 2003; **163**(3): 1147-51.
- 7 11. Sankoff, D. Minimal mutation trees of sequences. *SIAM J Appl Math. Siam Journal on Applied*
8 *Mathematics*. 1975; **28**(1): 35-42.
- 9 12. Hartigan, JA. Minimum mutation fits to a given tree. *Biometrics*. 1973; **29**(1): 53-65.
- 10 13. McCarthy, KR, Rennick, LJ, Nambulli, S, *et al.* Recurrent deletions in the SARS-CoV-2 spike
11 glycoprotein drive antibody escape. *Science*. 2021: eabf6950.
- 12 14. Kim, Y, Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining
13 chromosome. *Genetics*. 2002; **160**: 765-77.
- 14 15. Li, H, Stephan, W. Maximum likelihood methods for detecting recent positive selection and localizing
15 the selected site in the genome. *Genetics*. 2005; **171**: 377-84.
- 16 16. Bouckaert, R, Vaughan, TG, Barido-Sottani, J, *et al.* BEAST 2.5: An advanced software platform for
17 Bayesian evolutionary analysis. *PLoS Comput Biol*. 2019; **15**(4).
- 18 17. Gao, F, Ming, C, Hu, WJ, *et al.* New software for the fast estimation of population recombination rates
19 (FastEPRR) in the genomic era. *G3*. 2016; **6**(6): 1563-71.
- 20 18. Lam, HM, Ratmann, O, Boni, MF. Improved algorithmic complexity for the 3SEQ recombination
21 detection algorithm. *Mol Biol Evol*. 2018; **35**(1): 247-51.
- 22 19. Fu, Y-X. Statistical properties of segregating sites. *Theor Popul Biol*. 1995; **48**: 172-97.
- 23 20. Ohta, T, Kimura, M. On the constancy of the evolutionary rate in cistrons. *J Mol Evol*. 1971; **1**: 18-25.
- 24 21. Wang, Y, Dai, G, Gu, Z, *et al.* Accelerated evolution of an *Lhx2* enhancer shapes mammalian social
25 hierarchies. *Cell Res*. 2020; **30**(5): 408-20.
- 26 22. Fu, YX. Estimating mutation rate and generation time from longitudinal samples of DNA sequences.
27 *Mol Biol Evol*. 2001; **18**(4): 620-6.
- 28 23. Li, H, Stephan, W. Inferring the demographic history and rate of adaptive substitution in *Drosophila*.
29 *PLoS Genet*. 2006; **2**: e166.
- 30 24. Sagulenko, P, Puller, V, Neher, RA. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus*
31 *Evol*. 2018; **4**: vex042.
- 32 25. Li, Q, Guan, X, Wu, P, *et al.* Early transmission dynamics in Wuhan, China, of novel
33 coronavirus-infected pneumonia. *N Engl J Med*. 2020.
- 34 26. Hadfield, J, Megill, C, Bell, SM, *et al.* Nextstrain: real-time tracking of pathogen evolution.
35 *Bioinformatics*. 2018; **34**(23): 4121-3.
- 36 27. Flynn, JA, Purushotham, D, Choudhary, MNK, *et al.* Exploring the coronavirus pandemic with the
37 WashU Virus Genome Browser. *Nat Genet*. 2020; **52**: 986-1001.
- 38 28. White, WTJ, Holland, BR. Faster exact maximum parsimony search with XMP. *Bioinformatics*. 2011;
39 **27**(10): 1359-67.

- 1 29. Hendy, MD, Penny, D. Branch and bound algorithms to determine minimal evolutionary trees. *Math*
2 *Biosci.* 1982; **59**: 277-90.
- 3 30. Tang, X, Wu, C, Li, X, *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev.*
4 2020; **7**(6): 1012-23.
- 5 31. Forster, P, Forster, L, Renfrew, C, *et al.* Phylogenetic network analysis of SARS-CoV-2 genomes.
6 *Proceedings of the National Academy of Sciences of the United States of America.* 2020; **117**(17):
7 9241-3.
- 8 32. Qin, E, Zhu, QY, Yu, M, *et al.* A complete sequence and comparative analysis of a SARS-associated
9 virus (Isolate BJ01). *Chin Sci Bull.* 2003; **48**(10): 941-8.
- 10 33. Hu, B, Zeng, L-P, Yang, X-L, *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses
11 provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* 2017; **13**(11): e1006698.
- 12 34. Han, YL, Du, J, Su, HX, *et al.* Identification of diverse bat alphacoronaviruses and betacoronaviruses
13 in China provides new insights into the evolution and origin of coronavirus-related diseases. *Front*
14 *Microbiol.* 2019; **10**: 1900.
- 15 35. Rambaut, A, Loman, N, Pybus, O, *et al.* Preliminary genomic characterisation of an emergent
16 SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *virological.org.* 2020:
17 [https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-th](https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563)
18 [e-uk-defined-by-a-novel-set-of-spike-mutations/563](https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563).
- 19 36. Hartl, DL, Clark, AG. *Principles of Population Genetics.* Sunderland, Massachusetts: Sinauer
20 Associates, Inc.; 1988.
- 21 37. Li, J, Schneider, KA, Li, H. The hitchhiking effect of a strongly selected substitution in male germline
22 on neutral polymorphism in a monogamy population. *PLoS ONE.* 2013; **8**(8): e71497.
- 23 38. Stephan, W, Wiehe, THE, Lenz, MW. The effect of strongly selected substitutions on neutral
24 polymorphism: analytical results based on diffusion theory. *Theor Popul Biol.* 1992; **41**: 237-54.
- 25 39. Schraiber, JG, Evans, SN, Slatkin, M. Bayesian inference of natural selection from allele frequency
26 time series. *Genetics.* 2016; **203**(1): 493-511.
- 27 40. Kaplan, NL, Hudson, RR, Langley, CH. The "hitchhiking effect" revisited. *Genetics.* 1989; **123**:
28 887-99.
- 29 41. Korber, B, Fischer, WM, Gnanakaran, S, *et al.* Tracking changes in SARS-CoV-2 Spike: Evidence that
30 D614G increases infectivity of the COVID-19 virus. *Cell.* 2020; **182**(4): 812-27.
- 31 42. Plante, JA, Liu, Y, Liu, J, *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature.* 2020.
- 32 43. Turakhia, Y, Thornlow, B, Hinrichs, AS, *et al.* Ultrafast sample placement on existing trees (USHER)
33 empowers real-time phylogenetics for the SARS-CoV-2 pandemic. *BioRxiv.* 2020.
- 34 44. du Plessis, L, McCrone, JT, Zarebski, AE, *et al.* Establishment and lineage dynamics of the
35 SARS-CoV-2 epidemic in the UK. *Science.* 2021; **371**: 708-12.
- 36 45. Holshue, ML, DeBolt, C, Lindquist, S, *et al.* First Case of 2019 Novel Coronavirus in the United
37 States. *N Engl J Med.* 2020; **382**(10): 929-36.
- 38 46. Yu, D, Dong, L, Yan, F, *et al.* eGPS 1.0: comprehensive software for multi-omic and evolutionary
39 analyses. *Natl Sci Rev.* 2019; **6**: 867-9.