Supplemental methods and materials

Contents

Supplemental methods and materials

Data quality control

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

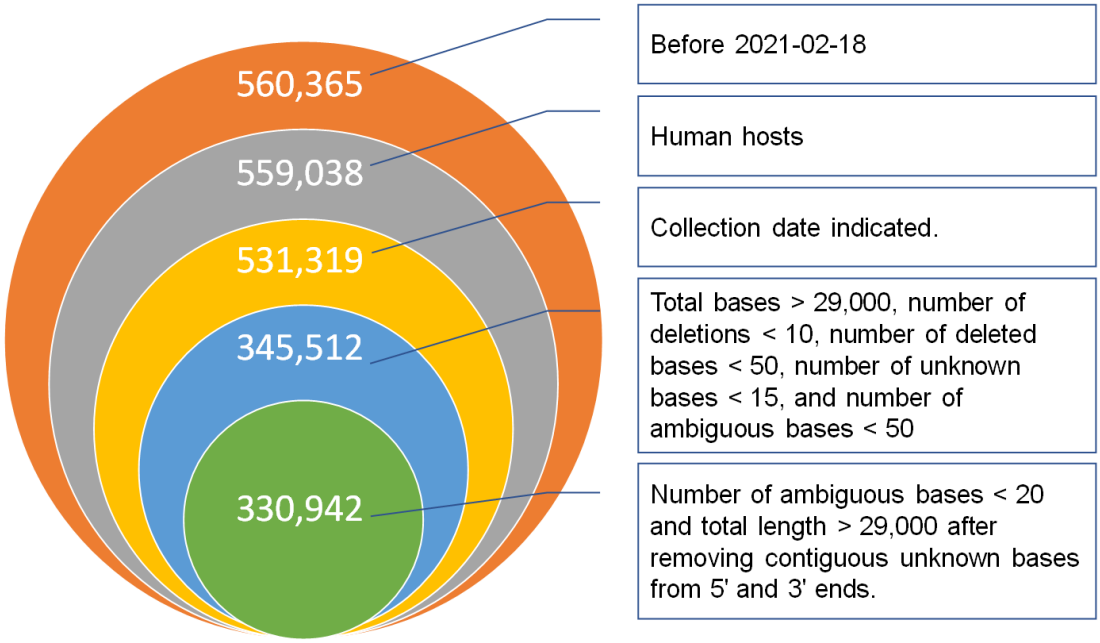

Figure S1. Quality control pipeline. The value in each circle is number of sequences identified in the quality control performed on February 18, 2021.

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

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

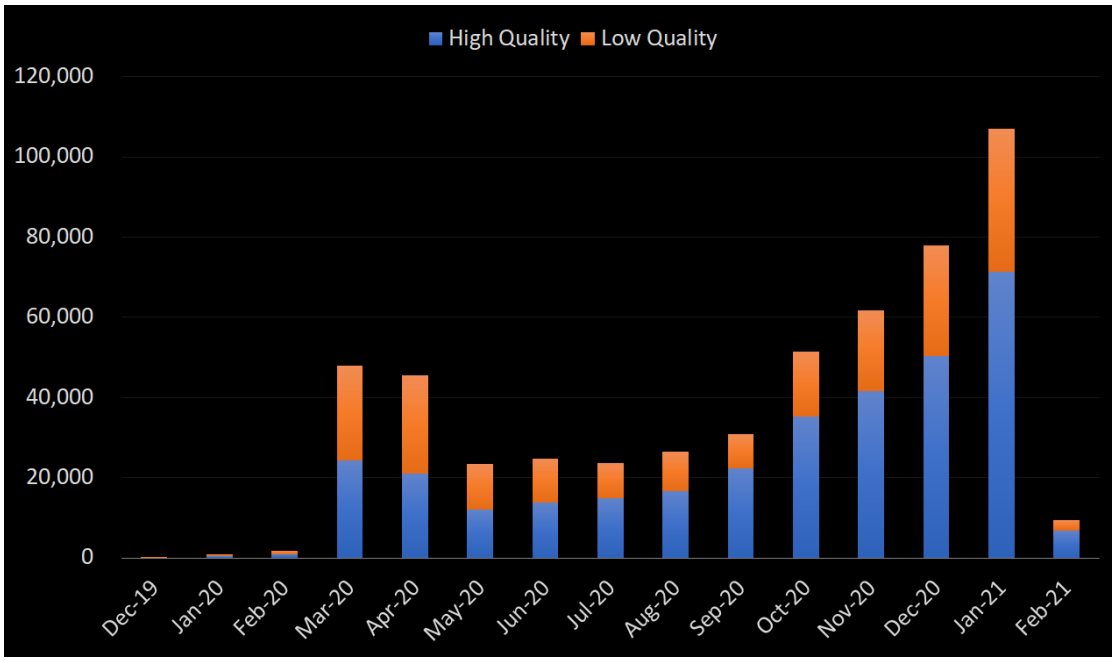

Figure S2. Number of high- and low-quality SARS-CoV-2 genomic sequences at various time points.

Distributed genome alignments

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

Ancestral alleles of SARS-CoV-2

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

Construction of the evolutionary tree based on distributed alignments

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

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

To solve this problem, the following strategy was used to treat ambiguous bases. For comparison of the sequence ACGRCG with the reference sequence ACGACG, ACGRCG is converted to ACGACG and ACGGCG. The resulting 2 new sequences are defined as one sequence set. Because this sequence set has the sequence ACGACG that is the same as that of the reference sequence, the strain with the sequence ACGRCG is considered as the same type as the strain with the reference sequence ACGACG. For comparison of the sequence ACGRCG with the sequence ACGYCG, ACGRCG is converted to ACGACG and ACGGCG, and ACGYCG is converted to ACGCCG and ACGTCG. Therefore, two sequence sets are generated. Because the 4 sequences in these two sequence sets are different, the strain with the sequence ACGRCG and the one with the sequence ACGYCG are considered as two different types. For comparison of the sequence ACGRCG with the sequence ACGHCG, ACGRCG is converted to ACGACG and ACGGCG, and ACGHCG is converted to ACGACG, ACGCCG and ACGTCG. As the resulting two sequence sets share the same sequence ACGACG, the strain with the sequence ACGRCG and the one with the sequence ACGHCG are considered as the same type.

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

Imputation of ambiguous and missing nucleotides

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

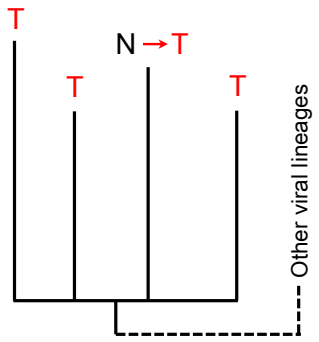

Figure S3. Imputation of ambiguous nucleotides of a lineage using the information of its siblings.

Parsimony inference of mutations for strains in each branch

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

Mutations affected by recombination

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

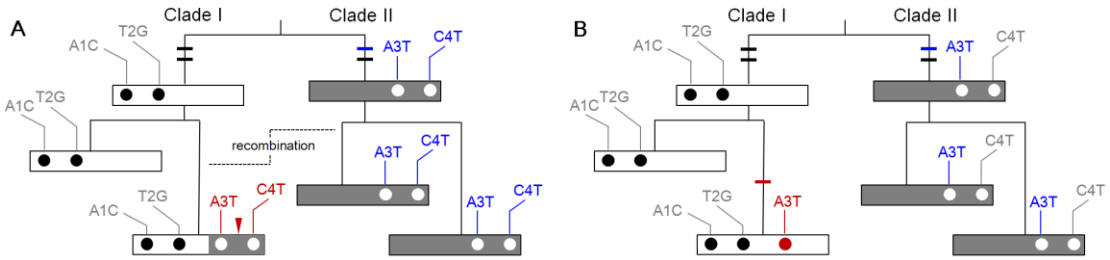

Figure S4. Generation of similar genomic variants by recombination and recurrent mutation.

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 [18]. 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 1 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.

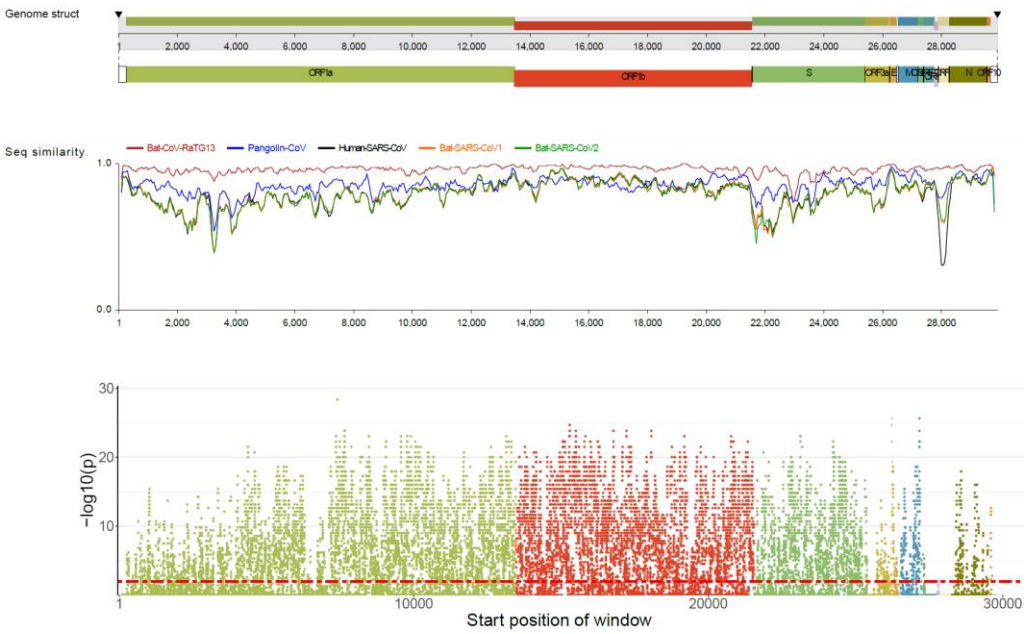

Figure S5. Manhattan plot of mutation cold spots in the genome of SARS-CoV-2. Results of genome-wide scan for mutation cold spots are shown in Manhattan plot of significance against SARS-CoV-2 reference genomic locations. In total, 330,942 high quality genomic sequences (submitted before February 18,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

To find mutation cold spots, the mutation density of a genome is denoted as β (mutations per base), and the observed number of mutations within a 10-base window is denoted as ξ_{obs}. Under the assumption of homogeneous mutation distribution, the expected number of mutations within the window is 10β. The significant level of mutation cold spots is determined by Poisson probability [19,20]: $P(x \leq \xi_{obs}) = \sum_{x \leq \xi_{obs}} e^{-10\beta} (10\beta)^x / x!$. It is a one-tailed test. Since a deletion may include multiple bases, the number of deleted bases, instead of the number of deletions, is used to determine the Poisson probability. If insertions are present, the window is ignored.

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

Estimation of mutation rate

Most SARS-CoV-2 strains were collected in different days. Similar to cases of longitudinal samples [21], more mutations are accumulated subsequent to the appearance of the most recent common ancestor (MRCA) [15].

Based on 178,765 high-quality genomic sequences submitted before January 5, 2021, linear regression was performed to estimate the mutation rate of SARS-CoV-2 (Figure S6). For each strain with a different collection date in a tip-dated time tree, the number of mutations, including that of recurrent mutations, was counted subsequent to the appearance of MRCA. As described previously, recombination-flagged mutations were excluded. Similar to the previous study [15], demography and the time of MRCA appearance were not required for estimation of mutation rates.

The regression line ($y = 0.0553x + 0.8086$) is obtained (Figure S6), where *x* is the number of days between December 1, 2019 and date of collection, and *y* is the number of mutations accumulated since the first appearance of MRCA of SARS-CoV-2. The slop of the regression line indicates the genome-wide mutation rate of SARS-CoV-2 (0.0553 per genome per day or 6.8017×10^{-4} per nucleotide per year). As the extended regression line crosses the *x*-axis at -14.6 days, the time of the first appearance of MRCA of SARS-CoV-2 was determined to be November 15, 2019. The mutation rate of each SARS-CoV-2 gene is shown in Table S1.

The 95% confidence interval of the estimated mutation rate is obtained via Monte-Carlo simulations. Given the estimated mutation rate, mutations are randomly generated along the evolutionary tree [22], and mutation rate is estimated by regression analysis. Then the empirical distribution of estimated mutation rate is obtained from 1,000 simulated data set.

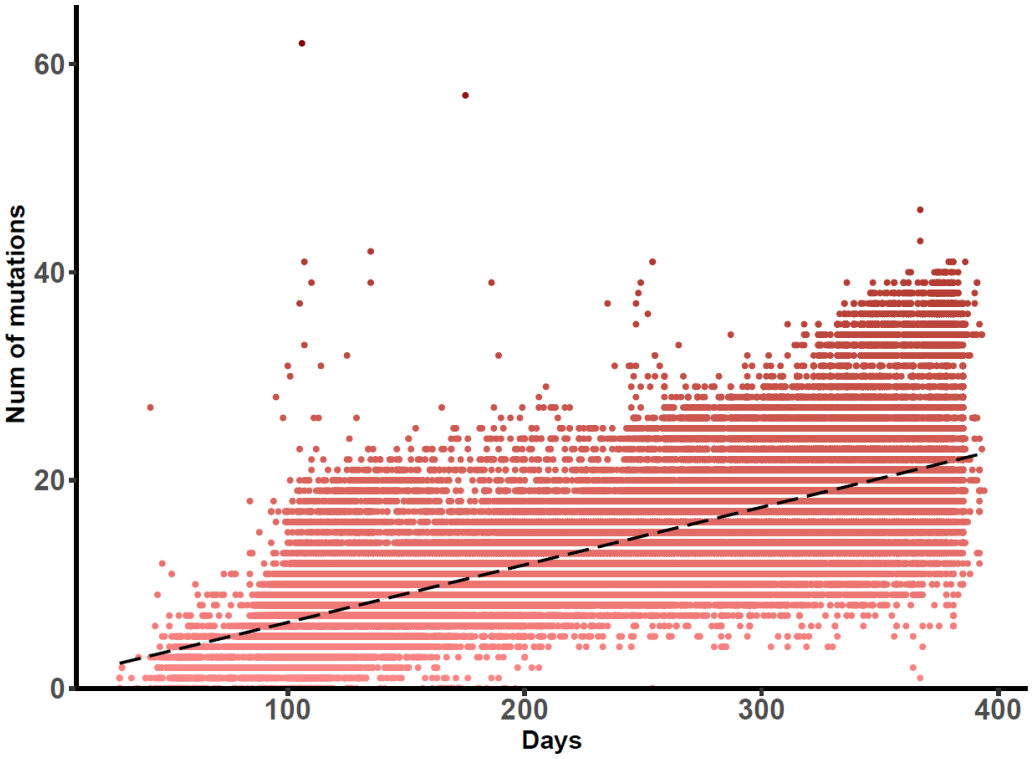

Figure S6. Linear regression for estimation of mutation rate. In this figure, 178,765 high-quality genomic sequences (submitted before January 5, 2021) were analyzed. The *x*-axis displays number of days between December 1, 2019 and date of collection. The *y*-axis indicates number of mutations accumulated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

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

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}

Maximum-likelihood phylodynamic analysis

A highly effective maximum-likelihood method (TreeTime) is used to determine the dates of internal nodes [23] as it allows fast inference by "the post- and pre-order traversals" with tabulated key values for back tracing. This algorithm was implemented in CGB with very minor revisions. The genome-wide mutation rate is also timely updated to calculate the likelihood.

As recommended by TreeTime [23], all length zero branches are pruned, and branch length corresponds to number of mutations on the branch. To improve computation efficiency, CGB first categories branches with length zero according to its context (Figure S7). In some cases, branches with length zero are not pruned (Figure S7A, E) in order to make length zero offspring as a clade and to reduce the number of multifurcated nodes.

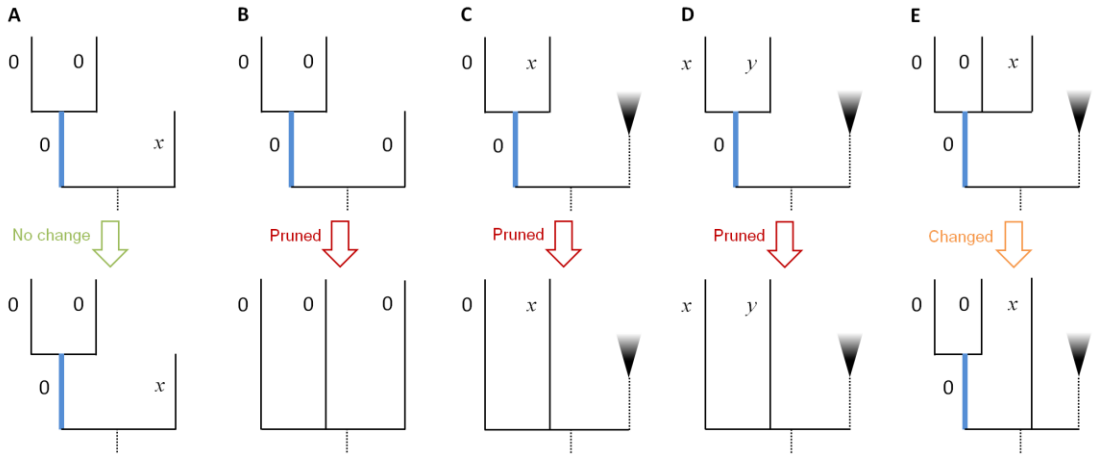

Figure S7. Five categories of length zero branches (highlighted in blue).
A) All offspring of the branch have length zero, and the sister branch of the branch has length non-zero *x*. In this case, the two offspring of length zero are in the same clade.
B) The sister branch has length zero, and the three nodes are clustered to form a multifurcated clade.
C) If one offspring of the branch has length zero, the branch is pruned.
D) If all offspring of the branch have length non-zero *x* or *y*, the branch is pruned.
E) If two or more offspring of the branch have length zero, the branch is kept and the non-zero branch is removed.

Many internal nodes are multi-furcated instead of bi-furcated because the viral strains are very similar to each other. The multi-furcated nodes are known as polytomies. To reduce the number of branches of a polytomic node, CGB sorts the branches according to the potential gain of likelihood if branches are shortened and determines whether a longer or shorter branch length would increase the likelihood of tree. The branches are bi-partitioned to form a new clade (Figure S8), and the two sets of branches are determined by maximizing the gain of likelihood. The bi-partition always starts from the root to the tips, and this process is repeated at least four times. The date of the MRCA appearance was estimated by this method to be November 15, 2019 (95% CI: October 21 – November 22, 2019), the same as that estimated by the regression analysis described above. As the first onset of COVID-19 was reported on December 8, 2019 [24], the result suggests that viral spread and evolution occurred before that date.

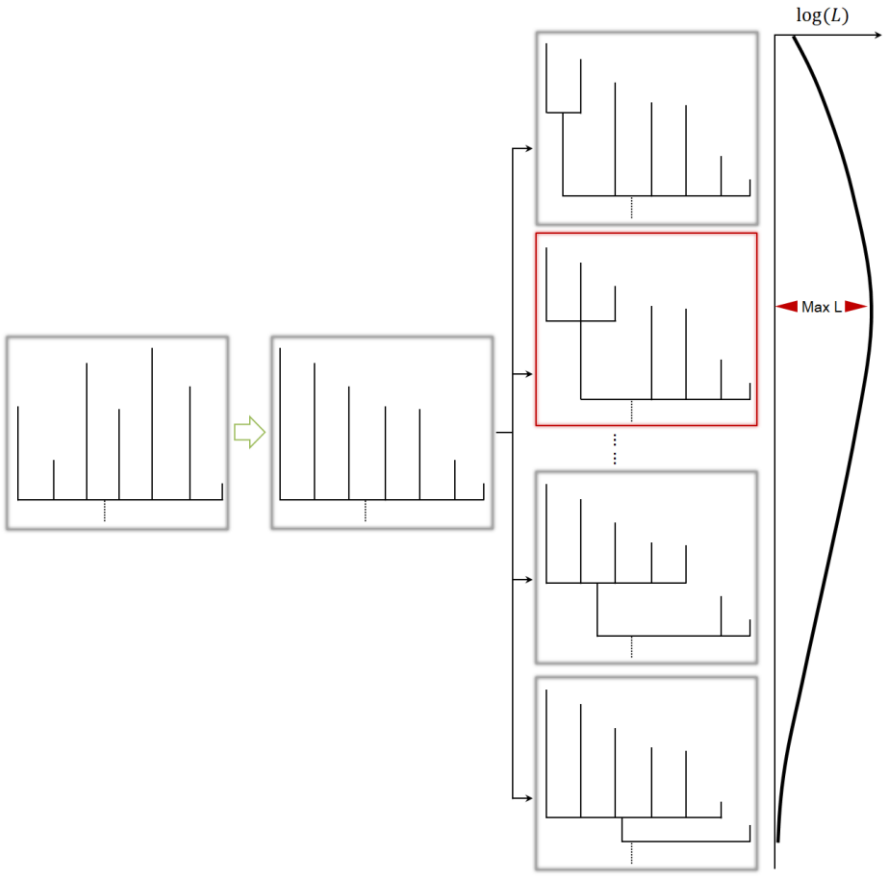

Figure S8. Bi-partition of a polytomic node. CGB first sorts the branches according to the potential gain of likelihood. If *k* branches are linked to the node, there are $k-2$ different ways to bi-partition the node. The two sets of branches are determined by maximizing the gain of likelihood.

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

Similar to NextStrain [25]and the WashU Virus Genome Browser [26], CGB uses a tree-based file format to show SARS-CoV-2 genomic variants. The head of data file contains data version, update date, genomic region analyzed, and mutation rate estimated for each gene. The data file is in Newick tree format (nwk) and contains information on collection date, gender and age of patient, location for each strain, mutations, and inferred internal nodes. Recombination flags are not included in the output data because they can be easily reconstructed. To allow fast showing and re-analyzing large number of SARS-CoV-2 genomic variants, redundant data are minimized.

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

Branch and bound for maximum parsimony [27,28] is implemented with a speed-up revision. New genomic sequences of SARS-CoV-2 strains are first aligned with the reference genome (Figure 1). The resulting alignment and previous results are then analyzed together, and the evolutionary tree is rebuilt using previous result file that contains the existing tree and mutation information. A new strain is then added to the mutation-annotated tree as a dated leaf, and new mutations are labeled and analyzed according to the principle of parsimony. CGB adds the earliest strain first to the tree. After adding all new genomic sequences, the mutation rate of SARS-CoV-2 is calculated, and the date of each internal node is determined as described above. This maximum-likelihood (ML) analysis was performed with a slightly revised version of TreeTime [23].

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

CGB binary nomenclature for each internal node or branch

A number of different naming systems have been proposed [25,29,30], but these systems only name a very small number of internal nodes or branches. As there are 98,496 internal nodes on the huge evolutionary tree of 330,942 SARS-CoV-2 strains, it is

nearly impossible to manually label each node. Therefore, the CGB binary nomenclature system was developed following the MRCA concept as follows.

Each node of a viral strain is first assigned a permanent unique positive integer (e.g., 1 – 9) in the order of discovery (Figure S9). Assuming that an internal node has 2 sub-nodes that are named CGB1 and CGB2, this internal node is named CGB1.2. For an internal node with more than 2 sub-nodes, e.g., CGB7, CGB9, and CGB6, it is named with the two smallest CGB numbers, given the condition that the internal node is the MRCA of the two sub-nodes, separated by a dot; thus, this internal node is designated as CGB6.7.

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

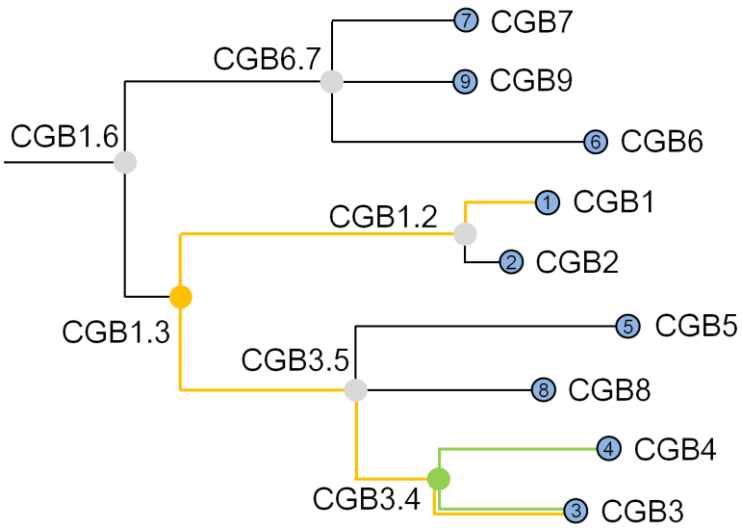

Figure S9. Illustration of CGB binary nomenclature. The evolutionary tree is shown with 9 strains named CGB1 – CGB9. The green internal node with two sub-nodes named CGB3 and CBG4 is designated as CGB3.4 since the MRCA of CGB3 and CGB4 is the green node. For an internal node with more than 2 sub-nodes, it is named with the two smallest CGB numbers, given the condition that the internal node is the MRCA of the two sub-nodes, separated by a dot. Therefore, an orange internal node is named CGB1.3 because it contains CGB1, CGB2, CGB5, CGB8, CGB4, and CGB3 with 1 and 3 being the smallest CGB numbers, on the condition just described.

Visualization of the huge evolutionary tree by movie-maker strategy

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

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

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

Tree visualization with CGB

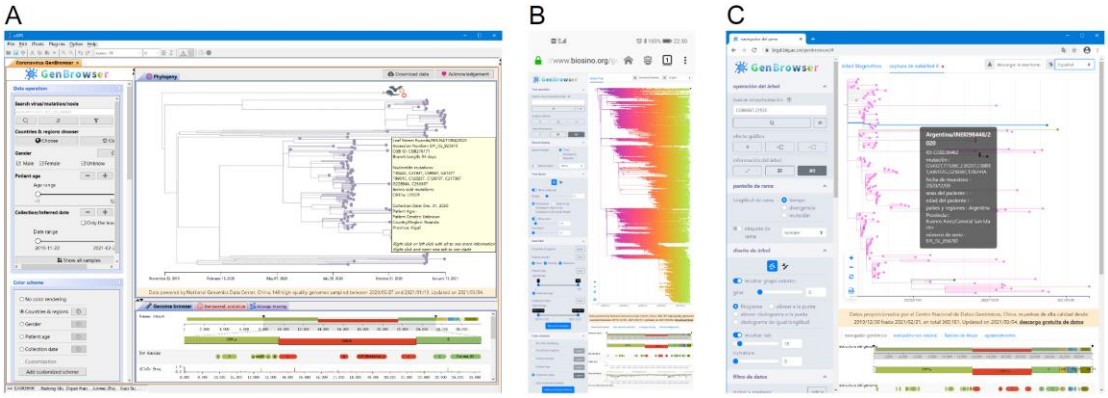

Figure S10. Tree visualization with CGB.

A) Tree visualization of 148 SARS-CoV-2 strains collected from Rwanda, Africa.

B) Web-based CGB tree visualization of 360,181 genomes with the Android version of Firefox.

C) Web-based CGB tree visualization of an Argentinian clade (CGB6867.22533) in Spanish with the desktop version of Google Chrome. Nine language versions (Chinese, English, German, Japanese, French, Italian, Portuguese, Russian, and Spanish) are available.

Coordinated annotation tracks

CGB uses six tracks to show genome structure and key domains; allele frequencies; sequence similarity between various coronavirus isolated from human, bat, and pangolin; multi-genome alignment; and primer sets for detection of various SARS-CoV-2 genes and strains (Figure S11). These tracks are coordinated according to nucleotide positions of the SARS-CoV-2 reference genome.

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

The third track shows the frequencies of derived alleles or variants (Figure S11). Since the web version is designed for the general public and quick view of global samples, users can update manually the frequency of an allele in the chosen clade.

When hovering mouse on the frequency column of an allele, its allele frequency trajectory (Figure S12) will pop up. This allele frequency trajectory is calculated by a sliding window of 5 days in size. The person who first discovered the allele is indicated below allele frequency trajectory.

Figure S11. Six tracks shown by the Coronavirus GenBrowser.

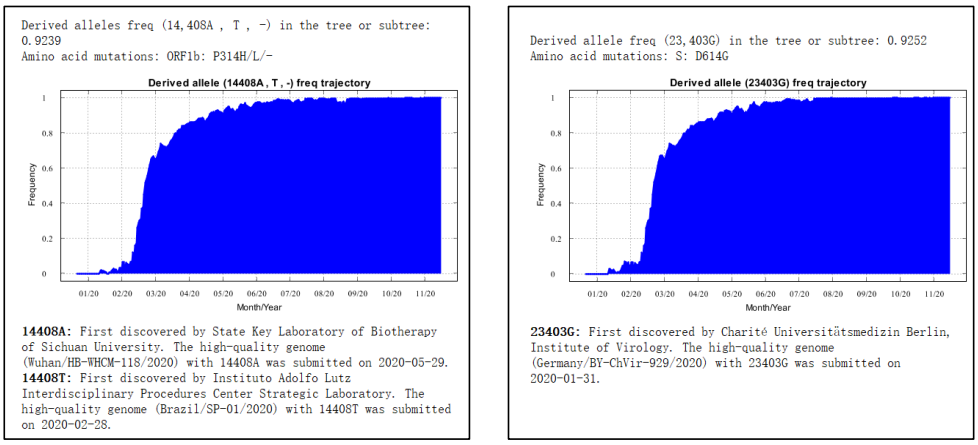

Figure S12. Visualization of allele frequency trajectory with CGB.

The fourth track shows sequence similarity between SARS-CoV-2 reference genome (NC_045512.2) [3] and the genomes of five other coronaviruses, including bat-CoV-RaTG13 (MN996532.1) [4], pangolin-CoV (MT040334.1) [5], human-SARS-CoV (AY278488.2) [31], bat-SARS-CoV1 (KY417146.1) [32], and bat-SARS-CoV2 (MK211376.1) [33]. Sequence similarity is determined using a sliding window (window size 100 bp and sliding step 20 bp). In the standalone version of CGB, these parameters can be adjusted to re-calculate the degree of sequence similarity.

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

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

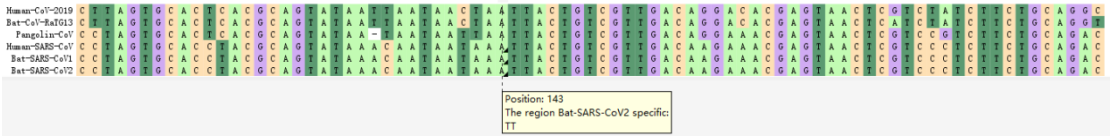

Figure S13. Multiple-genome alignment. Inserted sequences in the five non-SARS-CoV-2 genomes are marked with black triangles. This alignment can be downloaded from https://bigd.big.ac.cn/ncov/apis/.

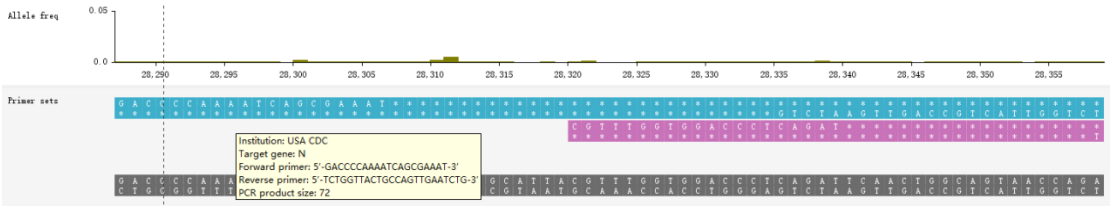

Figure S14. Combined view of two tracks of allele frequency and primer set for detection of various SARS-CoV-2 strains. The nucleotide sequences of two primers are shown, and their amplified region is marked in pink.

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

To detect branch-specific accelerated evolution, each internal branch of the SARS-CoV-2 tree was examined. For each internal branch, the observed number of mutations of the *i*-th gene ($\gamma_{obs,i}$) was compared with the expected number of mutations of the same gene ($\gamma_{exp,i}$). The significance level of acceleration was determined by Poisson probability [19,20]: $P(x \geq \gamma_{obs,i}) = \sum_{x \geq \gamma_{obs,i}} e^{-\gamma_{exp,i}} (\gamma_{exp,i})^x / x!$, where $\gamma_{exp,i} = t\mu_i l_i / 365$, t is the duration (in days) of the branch, μ_i is the

mutation rate (per nucleotide per year) of the *i*-th gene calculated from the entire collection of viral strains (Table S1), and l_i is the length of the *i*-th gene. It is a one-tailed test. The condition $t > 10$ (days) was used for detection of branch-specific accelerated evolution of SARS-CoV-2.

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

The internal branch CGB267525.267526 (genome-wide FDR corrected $P = 1.2302 \times 10^{-10}$) that had the most accelerated evolution was dated November 2020 with only two descendants. To determine whether variants with accelerated evolution were advantageous in spreading, the median number of descendants of other variants in the same time period (± 10 days) were determined and found to be 7. The probability for the CGB267525.267526 variant to have 2 or more than 2 descendants is 0.7762 (one-tailed). Thus the accelerated variant may not be more contagious than others.

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

Sequential occurrence of B.1.1.7-associated mutations

The variant of concern B.1.1.7 (CGB75056.269896) identified in the UK was recently reported [34]. The CGB evolutionary tree shows the sequential occurrence of B.1.1.7-associated non-synonymous mutations (Figure S15). The results indicate that S:HV69-, S:Y144-, and ORF1a:SGF3675- were recently occurred.

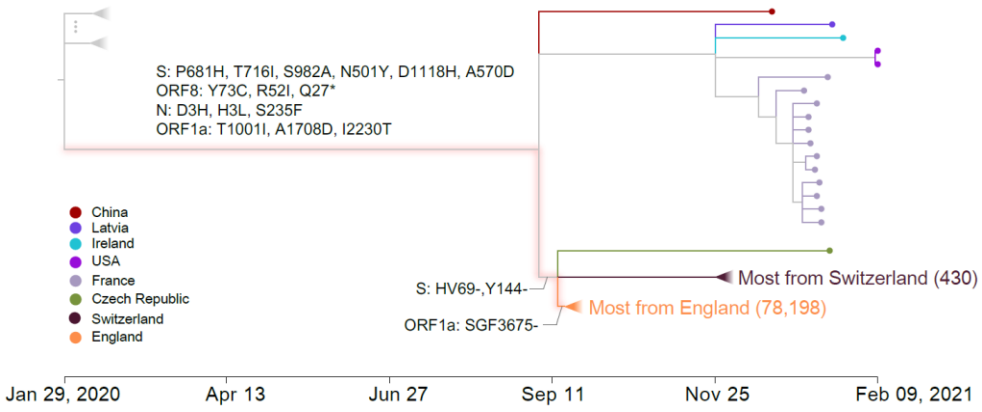

Figure S15. Tree visualization showing the sequential occurrence of B.1.1.7-associated non-synonymous mutations.

No correlation between ORF1ab and spike gene mutation rates

ORF1ab encodes a polyprotein that is involved in genome transcription and replication. Analysis was performed to determine whether the number of accumulated non-synonymous mutations in ORF1ab correlates with the mutation rate of spike gene. For each strain, the number of non-synonymous mutations in ORF1ab accumulated after the appearance of MRCA of SARS-CoV-2 was determined. The spike mutation rate was estimated based on the number of nucleotide mutations accumulated after the appearance of MRCA of SARS-CoV-2 divided by the length of duration (in days). No correlation between ORF1ab and spike gene mutation rates was observed (Pearson's correlation coefficient $= -0.00929$) (Figure S16).

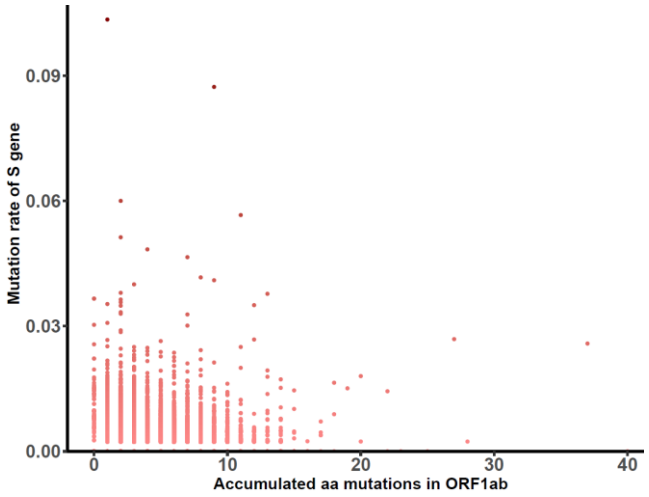

Figure S16. Correlation between number of ORF1ab non-synonymous mutations and spike gene mutation rate. The number of high quality SARS-CoV-2 genomic sequences is 330,942.

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

CGB can also determine whether a SARS-CoV-2 strain evolves more slowly recently. Based on 178,765 high-quality SARS-CoV-2 genomic sequences, the genome-wide mutation rate was determined to be $\mu = 6.8017 \times 10^{-4}$ per nucleotide per year. It indicates that a mutation occurs in a strain every 18.08 days. To detect variants with reduced evolution rate, CGB determines the number of days during which a chosen strain did not mutate. With Poisson probability [20], CGB also determines whether the number of mutations of the strain is fewer than that expected during the time period (one-tailed) (Figure S17). Detailed results of this analysis on three closely related internal branches are shown in Figure 4A.

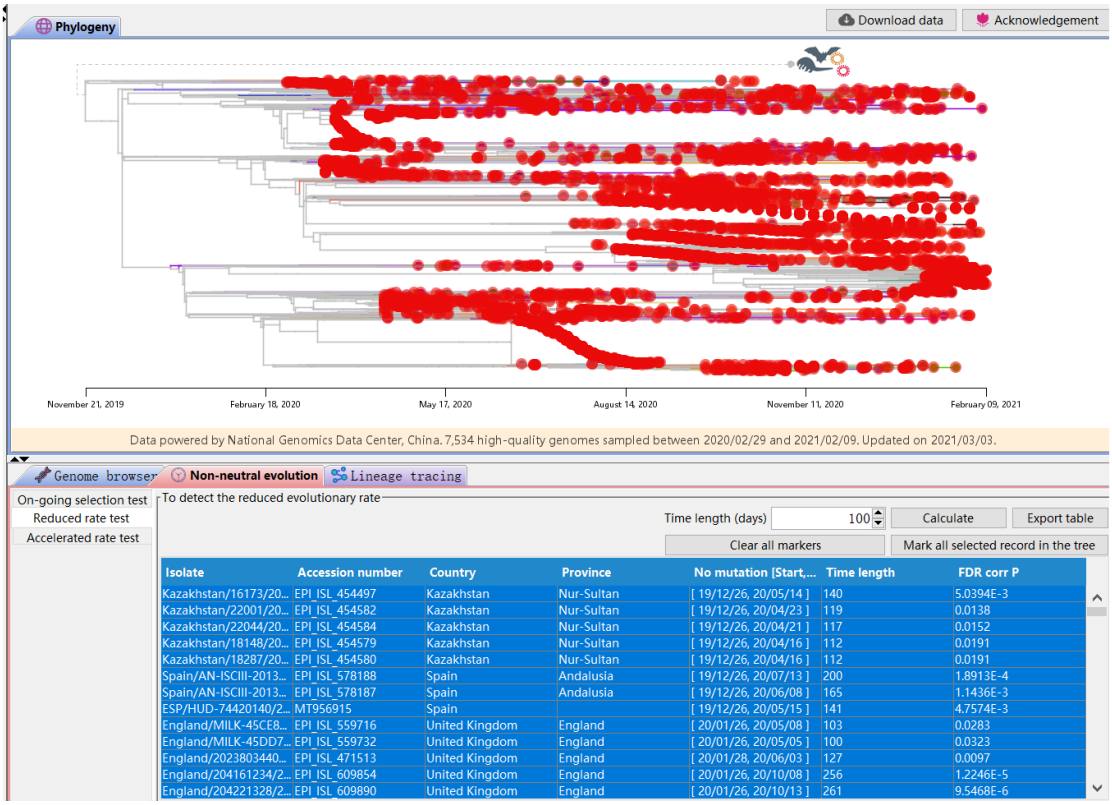

Figure S17. SARS-CoV-2 strains with reduced evolution rate among 330,942 SARS-CoV-2 genomic sequences. Each red dot represents a strain that evolved more slowly recently. The table shows some examples of such strains. The three analyzed closed related branches (Figure 4A) are marked with a blue empty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

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

To detect on-going positive selection, allele frequency trajectory with an S-shaped curve is examined as described previously [35-38]. For this determination, the selection coefficient is denoted as s. The initial frequency of the derived allele *a* is denoted as q_0, and that of the wide-type allele *A* is $p_0 = 1 - q_0$. The frequency of the wild-type allele *A* at a specific day (time *t*) is p_t, and that of the derived allele *a* is q_t.

The following equation was used to calculate the coefficient of on-going positive selection [35] (Table S2):

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

Then

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

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

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	$p_t = \frac{p_i}{p_i + (1+s)q_i}$	$q_t = \frac{(1+s)q_i}{p_i + (1+s)q_i}$

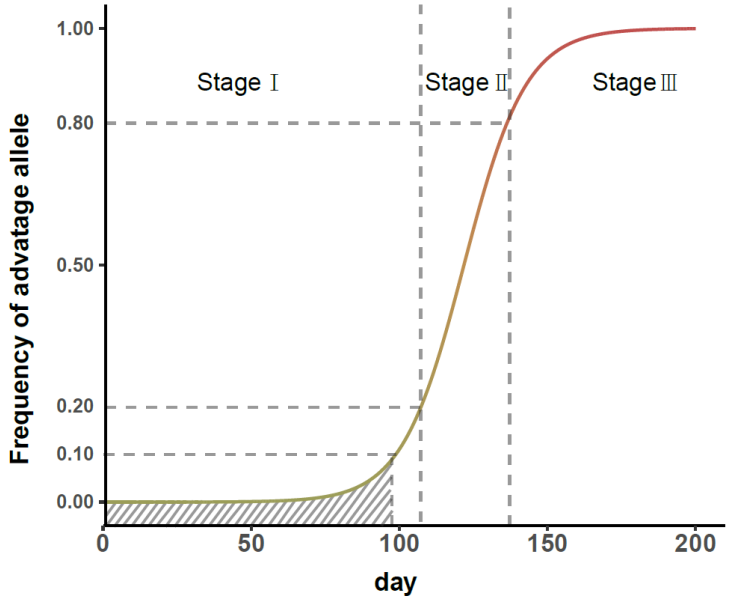

Figure S18. S-shaped frequency trajectory of advantageous mutations. $s = 0.1$ and $q_0 = 0.0001$. The best time window to control the transmission of a strain with an advantageous mutation is shadowed.

When $s > 0$, the frequency of a derived allele increases over time [35] (Figure S18). During Stages I and III, the speed of increase in the frequency of advantageous allele is slow, indicating low selection efficiency. During Stage II, the speed of increase in the frequency of advantageous allele is fast, and the efficiency of selection is high. When the frequency is 50%, the efficiency of selection reaches maximum. Therefore, the best time window to control the transmission of strains with an advantageous mutation is that of Stage I, especially when its frequency is still below 10%.

The analysis framework for detecting strains with putative advantageous mutations during their early stage of spreading is summarized in Figure S19. A neutral mutation may be linked to an advantageous mutation and spread over the entire population [13,14,39]. To reduce the impact of hitchhiking by neutral mutation, only non-synonymous mutations were analyzed. For this analysis, the initial (start) frequency must be < 0.1, and the end frequency must be > 0.05. Only the mutation frequency trajectory during the selective phase was used for calculation as this is the period when an advantageous mutation causes on-going selection. Linear regression analysis was performed to detect advantageous mutation. According to the equation described above, a mutation was considered advantageous when $s > 0$, $p < 0.01$, and $R^2 > 0.5$.

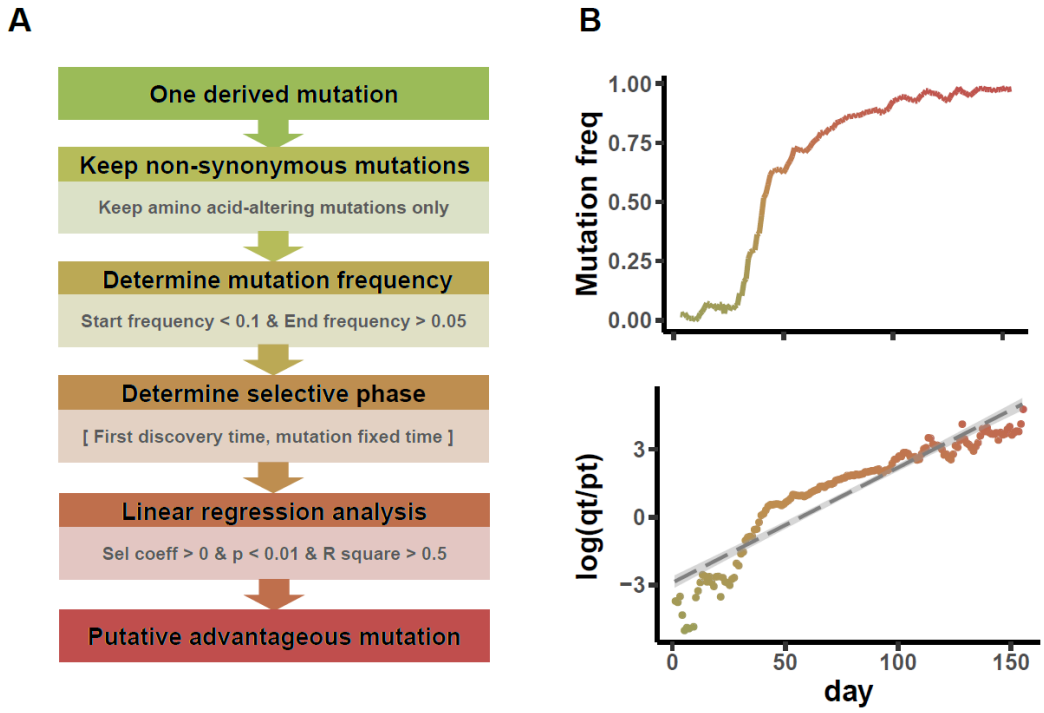

Figure S19. Detection of on-going selection of SARS-CoV-2.

A) Flow chart for detection of putative advantageous variants.

B) Frequency trajectory for A23403G (S: D614G) and linear regression analysis. The *x*-axis displays number of days since the first appearance of a derived allele in global virus population. q_t is the frequency of the derived allele (23403G), and p_t is the 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. [40,41]
*Mutations found on clade CGB75056.269896 (B.1.1.7).

Lineage tracing

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

With CGB, lineage tracing of the entire collection of 360,181 strains can be completed in seconds if aligned sequences are used as input. A similar approach (Ultrafast sample placement on existing trees, UShER) was independently developed recently [42] and published on bioRxiv on September, 28, 2020, while CGB lineage tracing was first released online (http://www.egps-software.net/) on August 12, 2020.

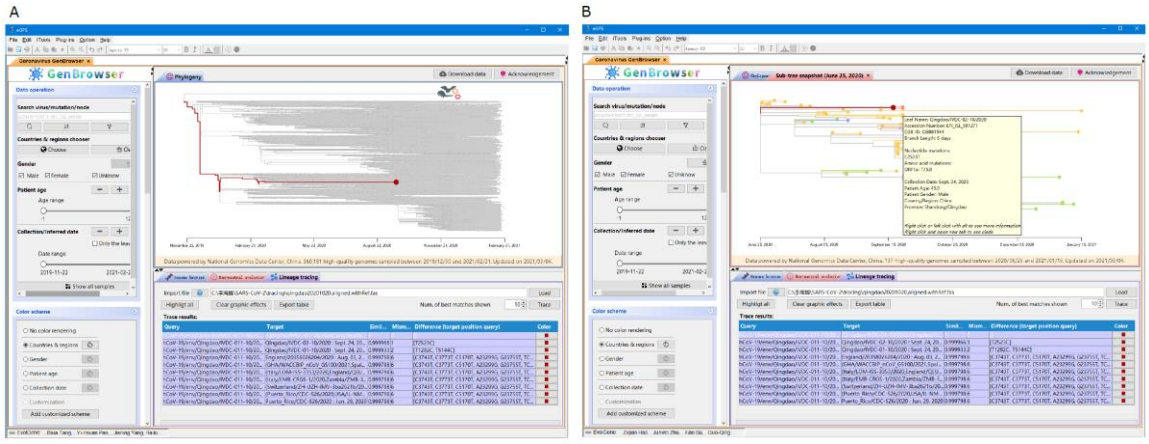

Figure S20. Lineage tracing with CGB.

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

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

Analysis of SARS-CoV-2 transmission in Washington State

The first outbreak of COVID-19 in Washington State was analyzed using the sequences of 453 SARS-CoV-2 isolates collected between February 20 and March 15, 2020 [6]. Results suggest that this outbreak was derived from a single introduction. Since then, more isolates were collected, and a total of 5,170 high-quality genomic sequences were obtained. Results of phylogenetic studies showed that these 5,170 strains belong to 4 clades (Figure S21). Therefore, deep sequencing could provide more details in the analysis of genomic epidemiology [43].

The CGB38.59 clade described before [6] contains 1,110 strains with the following mutations: C8782T and T28144C. The virus isolated from the first COVID19 case in the United States [44] is marked in red in Figure S21. This strain has single-nucleotide polymorphisms (SNPs) C8782T, C18060T, and T28144C compared to the MRCA of SARS-CoV-2. The CGB35.120 clade contains 4,033 strains. All strains in this clade have the following mutations: C241T, C3037T, C14408T, and A23403G. Strains in this clade are prevalent (4,033 / 5,170=78.0%) probably because they carry the advantageous mutation D614G (A23403G) in the spike protein [40]. Two minor clades (CGB1.7 and CGB58.127) are linked to the root of the tree. These clades are shown in the sub-panel (Figure S21) when major clades are collapsed.

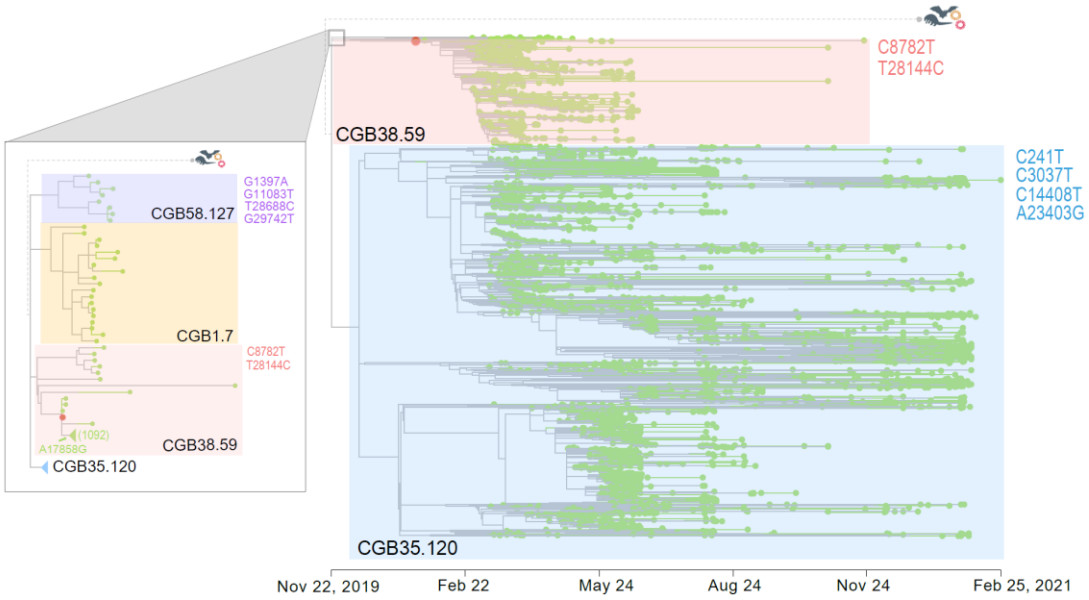

Figure S21. Phylogeny of 5,170 high-quality SARS-CoV-2 genomic sequences of strains found in Washington State.

The virus isolated from the first COVID19 case in the United States [44] is marked with a red circle (in CGB38.59 clade). This strain has single-nucleotide polymorphisms (SNPs) C8782T, C18060T, and T28144C compared to the MRCA of SARS-CoV-2. The mutations that distinguish these clades are indicated. The sub-panel shows enlarged view of the two minor clades.

Phylogenetic analysis of strains in the Auckland outbreak

There was an outbreak in Auckland in August 2020 after more than 100 days without a local transmission of COVID-19 in New Zealand. The first person who was tested positive in this outbreak worked at Americold food cold-storage facility in Auckland. CGB phylogenetic analysis revealed that all 56 strains in the CGB55430.55433 clade isolated from this outbreak carry a novel mutation T15867G, suggesting that the outbreak was derived from a single strain. This variant, as a genomic signature, was only found in strains associated with the Auckland outbreak (Figure S22).

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

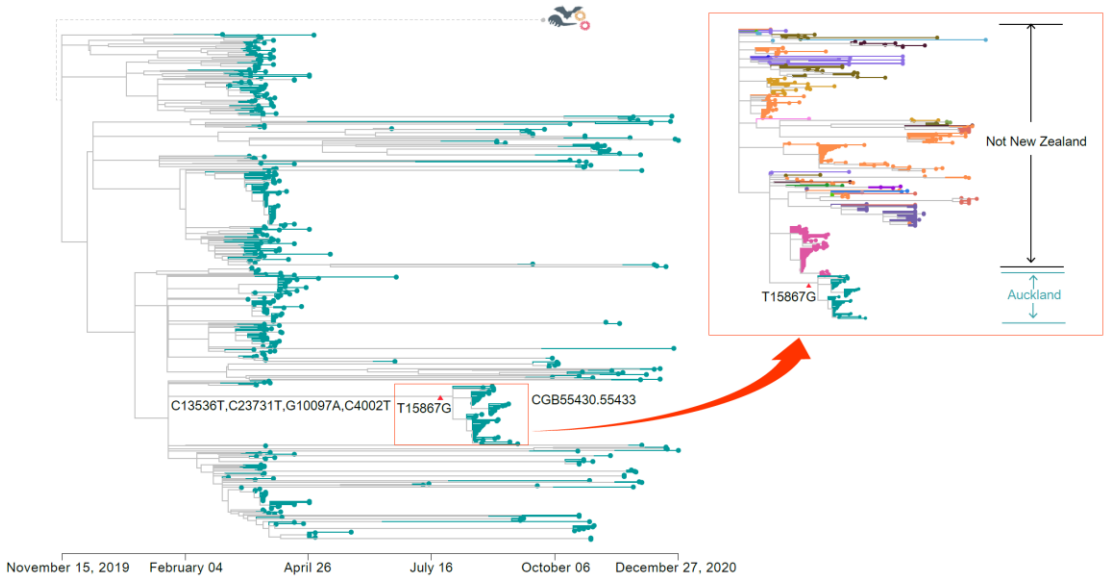

Figure S22. Phylogeny of 483 high-quality SARS-CoV-2 genomic sequences of strains found in New Zealand. The CGB55430.55433 clade of strains from the Auckland outbreak is highlighted. The mutations that distinguish strains of the Auckland outbreak from those of other outbreaks in New Zealand are indicated.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

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

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

D614G mutation in the spike protein

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

Standalone and web-based CGB

The web-based CGB is a simplified version in multiple languages and can be accessed with any web browser to view the tree, search a viral strain or a mutation, and perform data filtering. However, it does not monitor non-neutral evolutions (including accelerated and reduced evolution) and perform lineage tracing. It also does not have sufficient speed for viewing allele frequency trajectory and cannot zoom in to view an individual lineage. These functions are available in the standalone alone CGB. The standalone alone CGB is a plug-in module for the eGPS software (http://www.egps-software.net/) [45]. It provides the full function of CGB, and allele frequency trajectory can be promptly obtained.

CGB in multiple languages

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

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

Timely update of CGB

CGB provides timely or daily updates as needed.

Statistical information

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

Data availability

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

References

1. Zhao, W.-M. *et al.* The 2019 novel coronavirus resource. *Yi Chuan* **42**, 212-221 (2020).
2. Rozewicki, J., Li, S., Amada, K.M., Standley, D.M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res* **47**, W5-W10 (2019).
3. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).
4. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).
5. Lam, T.T.-Y. *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282-285 (2020).
6. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571-575 (2020).
7. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2-Approximately maximum-likelihood trees for large alignments. *PloS ONE* **5**, e9490 (2010).
8. Kozlov, A.M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast,

scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453-4455 (2019).

9. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-425 (1987).

10. Li, H., Zhang, Y.W., Zhang, Y.P. & Fu, Y.X. Neutrality tests using DNA polymorphism from multiple samples. *Genetics* **163**, 1147-1151 (2003).

11. Hartigan, J.A. Minimum mutation fits to a given tree. *Biometrics* **29**, 53-65 (1973).

12. McCarthy, K.R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science*, eabf6950 (2021).

13. Kim, Y. & Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765-777 (2002).

14. Li, H. & Stephan, W. Maximum likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* **171**, 377-384 (2005).

15. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**(2019).

16. Gao, F., Ming, C., Hu, W.J. & Li, H.P. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3* **6**, 1563-1571 (2016).

17. Lam, H.M., Ratmann, O. & Boni, M.F. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. *Mol Biol Evol* **35**, 247-251 (2018).

18. Fu, Y.-X. Statistical properties of segregating sites. *Theor Popul Biol* **48**, 172-197 (1995).

19. Ohta, T. & Kimura, M. On the constancy of the evolutionary rate in cistrons. *J Mol Evol* **1**, 18-25 (1971).

20. Wang, Y. *et al.* Accelerated evolution of an *Lhx2* enhancer shapes mammalian social hierarchies. *Cell Res* **30**, 408-420 (2020).

21. Fu, Y.X. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol Biol Evol* **18**, 620-626 (2001).

22. Li, H. & Stephan, W. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet.* **2**, e166 (2006).

23. Sagulenko, P., Puller, V. & Neher, R.A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042 (2018).

24. Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med* (2020).

25. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018).

26. Flynn, J.A. *et al.* Exploring the coronavirus pandemic with the WashU Virus Genome Browser. *Nat Genet* **52**, 986-1001 (2020).

27. White, W.T.J. & Holland, B.R. Faster exact maximum parsimony search with XMP. *Bioinformatics* **27**, 1359-1367 (2011).

28. Hendy, M.D. & Penny, D. Branch and bound algorithms to determine minimal evolutionary

trees. *Math Biosci* **59**, 277-290 (1982).
29. Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev* (2020).
30. Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci USA* **117**, 9241-9243 (2020).
31. Qin, E. *et al.* A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). *Chin Sci Bull* **48**, 941-948 (2003).
32. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog* **13**, e1006698 (2017).
33. Han, Y.L. *et al.* Identification of diverse bat alphacoronaviruses and betacoronaviruses in China provides new insights into the evolution and origin of coronavirus-related diseases. *Front Microbiol* **10**, 1900 (2019).
34. Rambaut, A. *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *virological.org*, 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 (2020).
35. Hartl, D.L. & Clark, A.G. *Principles of Population Genetics.*, (Sinauer Associates, Inc., Sunderland, Massachusetts, 1988).
36. Li, J., Schneider, K.A. & Li, H. The hitchhiking effect of a strongly selected substitution in male germline on neutral polymorphism in a monogamy population. *PLoS ONE* **8**, e71497 (2013).
37. Stephan, W., Wiehe, T.H.E. & Lenz, M.W. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor. Popul. Biol.* **41**, 237-254 (1992).
38. Schraiber, J.G., Evans, S.N. & Slatkin, M. Bayesian inference of natural selection from allele frequency time series. *Genetics* **203**, 493-511 (2016).
39. Kaplan, N.L., Hudson, R.R. & Langley, C.H. The "hitchhiking effect" revisited. *Genetics* **123**, 887-899 (1989).
40. Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812-827 (2020).
41. Plante, J.A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* (2020).
42. Turakhia, Y. *et al.* Ultrafast sample placement on existing trees (UShER) empowers real-time phylogenetics for the SARS-CoV-2 pandemic. *BioRxiv* (2020).
43. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708-712 (2021).
44. Holshue, M.L. *et al.* First Case of 2019 Novel Coronavirus in the United States. *N Engl J Med* **382**, 929-936 (2020).
45. Yu, D. *et al.* eGPS 1.0: comprehensive software for multi-omic and evolutionary analyses. *Natl Sci Rev* **6**, 867-869 (2019).