Evolutionary insights into a non-coding deletion of SARS-CoV-2 B.1.1.7

Jianing Yang[1,†], Guoqing Zhang[1,†], Dalang Yu[1,†], Ruifang Cao[1,†], Xiaoxian Wu[2], Yunchao Ling[1], Yi-Hsuan Pan[7], Chunyan Yi[3], Xiaoyu Sun[3], Bing Sun[3], Yu Zhang[2], Guo-Ping Zhao[1,2,5,*], Yixue Li[1,6,*], Haipeng Li[1,4,*]

[1]Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China.
[2]Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai 200032, China.
[3]Laboratory of Cell Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China.
[4]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China.
[5]School of Life and Health Sciences, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China.
[6]Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou 510005, China.
[7]Key Laboratory of Brain Functional Genomics of Ministry of Education, School of Life Science, East China Normal University, Shanghai 200062, China.

[†]These authors contributed equally.

[*]Corresponding authors: gpzhao@sibs.ac.cn; yxli@sibs.ac.cn; lihaipeng@picb.ac.cn

Abstract

The SARS-CoV-2 variant of concern B.1.1.7 has quickly spread. To identify its crucial mutations, we explored the B.1.1.7 associated mutations on an evolutionary tree by the Coronavirus GenBrowser and VENAS. We found that a non-coding deletion g.a28271-, at upstream of the nucleocapsid (*N*) gene, has triggered the high transmissibility of B.1.1.7. The deletion changes the core Kozak site of the *N* gene and may reduce the expression of N protein and increase that of ORF9b. The expression of ORF9b is also regulated by another mutation (g.gat28280cta) that mutates the core Kozak sites of the *ORF9b* gene. If both mutations back-mutate, the B.1.1.7 variant loses its high transmissibility. Moreover, the deletion may interact with ORF1a:p.SGF3675-, S:p.P681H, and S:p.T716I to increase the viral transmissibility. Overall, these results demonstrate the importance of the non-coding deletion and provide evolutionary insight into the crucial mutations of B.1.1.7.

Keywords

SARS-CoV-2; B.1.1.7; a28271-; Kozak site; ORF9b; N gene

SARS-CoV-2 lineage B.1.1.7, also known as Variant of Concern 202012/01, is a variant first detected in the UK in September 2020 (1) and have been established to be more transmissible than preexisting variants (2). Comparing with the reference genomic sequence of SARS-CoV-2 (Genbank Accession: NC_045512.2; GISAID ID: EPI_ISL_402125) (3), the B.1.1.7 carries 20 non-synonymous mutations and deletions in *ORF1ab*, spike (*S*), *ORF8*, and nucleocapsid (*N*) genes (Supplementary Table 1). There are no mutations occurred in the *ORF9b* gene, a complete internal ORF within the *N* gene. The frequencies of B.1.1.7 associated mutations reached quickly up to 76.92% (4). The B.1.1.7 strains not only have high transmissibility (2), quickly spread to many countries (5-8), but also increase risk of death (9, 10). It has been documented that S:p.N501Y (11, 12) and S:p.D614G (13, 14) increase the viral transmissibility, and S:p.P681H occurs on the spike S1/S2 cleavage site (15). All of the three mutations are carried by B.1.1.7 strains. However, its crucial mutations still remain unclear, and current studies focus on non-synonymous mutations, especially those occurred on the spike gene (16-18). In this study, we conducted a comprehensive evolutionary analysis. We show the multiple evidences that a non-coding deletion is a necessary condition for the high transmissibility of B.1.1.7.

A non-coding deletion triggered the high transmissibility of B.1.1.7

We examined the B.1.1.7 lineage using the Coronavirus GenBrowser (CGB) (4). The CGB evolutionary tree shows the sequential occurrence of B.1.1.7 characteristic mutations (Figure 1A). The results indicate that the deletion S:p.Y144- first appears and the mutation S:p.P681H follows. Then four branches emerge in the following order: the first branch with seven mutations (ORF1a:p.T1001I, p.A1708D, p.I2230T; S:p.T716I, p.S982A, p.D1118H; ORF8:p.R52I), the second with two mutations (S:p.A570D; ORF8:p.Y73C), the third with three mutations (N:p.D3H, p.H3L, p.S235F), and the forth with two mutations (S:p.N501Y; ORF8:p.Q27*). At last two deletions sequentially occur (S:p.HV69- and ORF1a:p.SGF3675-). Strikingly, we identify a non-coding deletion (g.a28271-) occurred right before the rapid spread of B.1.1.7. The deletion locates between *ORF8* and *N* genes. A dramatic difference is observed when comparing the number of strains with the deletion (B.1.1.7) and that without the deletion (B.1.1.7-like) (92,556 *vs* 259) (Figure 1A). Please note that B.1.1.7-like strains carry all the B.1.1.7 defining 20 non-synonymous mutations and deletions (Supplementary Table 1), including three beneficial mutations (S:p.N501Y, p.D614G, and p.P681H), but these strains do not have the high transmissibility. Therefore, the deletion g.a28271- may contribute markedly to the increase of B.1.1.7 transmissibility.

Pooling data from different countries/regions may create the problem of biased sampling due to sequencing capacity, and these countries may have different

anti-contagion policies on the COVID-19 pandemic (19). Thus, to avoid these effects, nine different regional data were used to re-examine the role of the non-coding deletion (Figure 2). The numbers of B.1.1.7 and B.1.1.7-like strains were obtained for England (76871 *vs* 27), Spain (712 *vs* 30), Switzerland (1332 *vs* 8), Germany (570 *vs* 2), USA (1028 *vs* 8), Australia (58 *vs* 1), South America (22 *vs* 1), Africa (86 *vs* 1), and Asia (642 *vs* 3). These numbers indicate an unbalanced transmissibility of strains with or without the non-coding deletion (g.a28271-). Therefore, the deletion g.a28271- has triggered the high transmissibility of B.1.1.7.

To confirm the evolutionary path of B.1.1.7, we applied VENAS (20) to obtain an evolution network of SARS-CoV-2 major haplotypes (Figure 1B). The network is divided into two clades L and S (21). The differences between the L/S clades are two viral genomic mutations g.c8782t and g.t28144c. The L clade is the major clade and can be further divided into four subclades, L1, L2, L3, and L4 (Figure 1B). The L1 subclade is characterized by three nonsynonymous mutations, ORF1a:p.P4715L, S:p.D614G, and N:p.RG203KR. The B.1.1.7 lineage is derived from the L1 subclade and evolves through the evolutionary phase3 and phase4 (Figure 1B): The phase3 feature mutation is S:p.Y144- (g.tta21991-). The mutation was first identified in India on 27 January, 2020, and then was detected in Italy on 19 March, 2020. The phase4 includes 16 non-synonymous variations (Supplementary Table 1, Figure 1B). The feature mutations were first identified in a UK strain (England/MILK-9E05B3/2020) collected on 20 September 2020. Then the deletion g.a28271- occurred. Therefore, the occurrence order of B.1.1.7 mutations in network-based analysis is in accordance with that in tree-based analysis.

The g.a28271- and g.gat28280cta changed the core Kozak sites in B.1.1.7

The non-coding deletion (g.a28271-) contributes markedly to the high transmissibility of B.1.1.7 (Figure 1A, and 2). The base 28,271 is located at the third base upstream of the start codon of the *N* gene. The deletion makes t28,270 to slip one base and changes the Kozak context of gene *N* from a suboptimal one (A at -3, T at +4) to an undesirable one (T at -3, T at +4) (Figure 3) (22). Based on the SARS sequence, Xu et al. (23) mutated the homological site to another undesirable one (C at -3, T at +4), the mutant reduces the expression of N protein and increases the translation of ORF9b protein. The ORF9b protein is translated via a leaky ribosomal scanning mechanism (23). Therefore, it is expected that the weak Kozak context affects the translational efficiency of gene *N* and increases the expression of *ORF9b* in B.1.1.7.

We also found that another B.1.1.7 mutation g.gat28280cta (N:p.D3L) changes the Kozak core sequence of *ORF9b* (22). The switch from a28281 to t28281 changes the Kozak context of *ORF9b* from an optimal context (A at -3, G at +4) to a suboptimal context (T at -3, G at +4) (Figure 3). It is expected that the expression

level of *ORF9b* protein may be affected or reduced (22). However, this remains to be determined with functional assay. Overall, these two mutations change the core Kozak sites and may affect the translational efficiency of gene *N* and *ORF9b*.

Reduced B.1.1.7 transmissibility associated with loss of g.a28271- and g.gat28280cta

Since certain B.1.1.7 strains have lost their characteristic mutations probably due to back mutations or recombination, we explored whether the loss of B.1.1.7 characteristic mutations would have an effect on the cumulative frequency of the mutated strains. We hypothesize that if the key mutations are lost, the cumulative frequency of the haplotype will be different with Hap1, the B.1.1.7 major type, and the reduced transmissibility can be detected. We compiled the B.1.1.7 strains into different haplotypes, according to the considered mutations (Figure 4A). In total, we observed 244 haplotypes defined by the B.1.1.7 characteristic mutations. We analyzed 12 haplotypes with $n \geq 40$. Among those haplotypes, Hap7 that carries the back mutations of g.a28271- and g.gat28280cta shows the largest deviation to the cumulative frequency of Hap1 (correlation coefficient, 0.2456) (Figure 4B, Supplementary Table 2), indicating the important role of this combination. Besides, the loss of one of the two mutations (Hap10 and Hap2) causes a relatively weak deviation. These pieces of evidences suggest that g.a28271- and g.gat28280cta are essential to the transmissibility of B.1.1.7.

High viral transmissibility associated with multiple B.1.1.7 mutations

We then examined whether the B.1.1.7 characteristic mutations are recurrent on the evolutionary tree of SARS-CoV-2. We used the CGB (4) to check their recurrent patterns. All of those mutations, including g.a28271- and g.gat28280cta (N:p.D3L), appeared multiple times in the evolutionary tree with 400,051 high-quality SARS-CoV-2 genomic sequences (Figure 5). Those recurrent mutations provide us great chances to explore their potential functions further.

To determine whether a variant carried a single B.1.1.7 mutation is advantageous in the viral transmission, the number of its descendants was compared with co-circulating variants (as control) during the same time period (Supplemental excel file). To ensure the power of this method, the predominate mutation S:p.D614G which indicates a fitness advantage (13) was used as an internal reference. Comparing with control, the variant (CGB35.11774, according to CGB binary nomenclature) with S:p.D614G clearly shows great transmission ability (with 38,7486 *vs* 9 descendants; *P*-value $= 4.88 \times 10^{-4}$). In total of 835 variants with a single B.1.1.7 mutation considered, none of them spread significantly faster than other variants in control, *i.e.*, no variants with a single B.1.1.7 mutation are found to have significant advantage in

transmission. This indicates that the rapid transmission of B.1.1.7 may require the interaction of multiple mutations.

To examine the combination effects among mutations, including direct or indirect interactions, we searched the variants that carry the crucial mutation g.a28271- and other B.1.1.7 characteristic mutations in non-B.1.1.7 clades. We found two clades as the evidences of interaction among mutations (Figure 6). For the first clade (CGB182694.380595), the mutation ORF1a:p.SGF3675- first occurs, and we do not observe a rapid spread until the second mutation g.a28271- occurs. The double-mutated variant appears to spread faster than its siblings, and the number of descendants is 618 and 21, respectively. This indicates that the combination of g.a28271- and ORF1a:p.SGF3675- may enhance the transmission of SARS-CoV-2.

Similarly, the two mutations (S:p.P681H, and S:p.T716I) first appear in the clade CGB199165.262639. Then the mutation g.a28271- occurs and the variant also shows a faster spread than its siblings. The number of descendants is 157 and 20, respectively. The mutation S:p.P681H is next to the furin cleavage site (15). The observation suggests that the mutation g.a28271- may interact with one, or both, of the mutations (S:p.P681H, and S:p.T716I).

Discussion

In this study, we conduct a comprehensive evolutionary analysis for B.1.1.7 characteristic mutations. Based on the evolutionary tree, the occurrence order of the mutations on the B.1.1.7 lineage and their recurrence on the non-B.1.1.7 lineages provides us the insight into the importance of the mutations. Then the non-coding deletion g.a28271- and its interacting mutations (g.gat28280cta, ORF1a:p.SGF3675-, S:p.P681H, and S:p.T716I) are identified. The mutations (g.a28271-, and g.gat28280cta) alter the Kozak sites of *N* and *ORF9b*, and may influence the translational efficiency of the two genes. Protein N has the highest translation rate (24, 25) and its expression is associated with the replication of the genomic RNA (26). The product of *ORF9b* has an interferon (IFN) antagonistic activity and can suppress the IFN production (27). So we suppose that the two mutations may change the expression level of the genes *N* and *ORF9b*, and then influence the viral transmission.

The sequence with accession EPI_ISL_601443 is the canonical B.1.1.7 VOC genomic sequence (28). However, it does not carry the key non-coding deletion g.a28271-. The deletion neither is a pre-identified genomic characteristic of the VOC, nor appears in the pathogen genomics platform Nextstrain (29). Therefore, we suggest using the sequence with accession EPI_ISL_629703 as the canonical B.1.1.7 VOC genomic sequence (collected 21 October, 2020, in UK) (Supplementary Figure 1).

Interestingly, it is very likely that g.a28271- occurs due to recurrent mutation instead of recombination in the B.1.1.7 lineage. First, the probability that g.a28271-

occurs should be high. There are four continuous ‘A’ nucleotides between 28,271 and 28,274. When one of these nucleotides is deleted, it causes the same effect. All of those deletions are categories as g.a28271- in CGB. Thus the deletion rate is quadrupled. Second, there is only one mutation g.a28271- on the identified branch (CGB84017.91425). Recombination tends to create a hybrid genomic structure (4, 30-33). The two previously mutated alleles (g28111, cta28280) remain unchanged when the mutation g.a28271- occurs although both mutated alleles are very close to the genomic position 28,271. Therefore, g.a28271- may have occurred as recurrent mutation in the B.1.1.7 clade.

ORF1a:p.SGF3675- located in the nsp6 may have effect on the formation of DMVs (double–membrane vesicles) which is the central hubs for viral RNA synthesis (15, 34) and S:p.P681H in the furin cleavage site is a known location of biological significance (1, 15). The coaction of these mutations may increase greatly the transmission ability of B.1.1.7. Therefore, our results provide insights into the relevant importance of B.1.1.7 characteristic mutations. However, we still have no clues about how three amino acid changes could have joint effects with the putatively altered expression level of the genes *N* and *ORF9b*. Please be aware that our analysis cannot guarantee to identify all key mutations, such as S:p.N501Y, S:p.D614G, and S:p.P681H (11, 13, 15).

Even though this study may not discover all the crucial mutations of B.1.1.7, our results indicate that the transmissibility of SARS-CoV-2 is highly variable. First, the viral transmissibility is not only determined by mutations of the *S* gene but also those located in other genes. Second, not only non-synonymous mutations but also non-coding one is important for the viral transmissibility. Third, it has been shown the acquisition of B.1.1.7 high transmissibility is the interactions of g.a28271- and multiple B.1.1.7 associated mutations. The beneficial B.1.1.7 haplotype may have occurred by chance due to a large number of combinations and recurrent mutations. This is different with the previously observed cases, such as S:p.D614G (13) which a single mutation can be beneficial. Therefore, considering that SARS-CoV-2 is mutating and new strains are forming, it is essential to establish convenient genomic surveillance platforms, make bioinformatics analysis easier in local agencies of public health, and discover the emergence of beneficial haplotypes in time.

Methods

Identification of recurrent mutations in the CGB evolutionary tree

Considering the degeneracy of the genetic code, we searched recurrent amino acid mutations by using the form of amino acid change, instead of that of nucleotide change in the CGB. To search g.a28271- in the evolutionary tree, we used the string “A28271-”.

Detection of transmission advantage for variants with single B.1.1.7 mutation

We first searched all the nodes that carry a B.1.1.7 mutation. For each node, its cumulative mutations are examined. If multiple B.1.1.7 mutations are found, the node is discarded. We only considered the nodes with single B.1.1.7 mutation to detect their transmission advantage. For each considered node (named as target node), we collected all other nodes dated in the same time period (±10 days). If less than 200 nodes were found, the considered time period would be expanded. The numbers of their descendants were obtained and used as the empirical distribution of descendant number. The parent node and child nodes of the target node were excluded to avoid the potential biased distribution. Then the descendant number of the target node was obtained and compared with the empirical distribution to calculate the *P*-value (one-tailed). Since we tested the advantageous of single B.1.1.7 mutation, descendants carrying multiple B.1.1.7 mutations were not counted.

Haplotype analysis

The B.1.1.7 clade (CGB84017.91425) was invoked to employ haplotype analysis. Considered genomic positions (3,267; 5,388; 6,954; 11,288-1,1296; 21,765-21,770; 21,991-21,993; 23,063; 23,271; 23,604; 23,709; 24,506; 24,914; 27,972; 28,048; 28,111; 28,271; 28,280-28,287; 28,977) were used to determine haplotypes. The sizes of haplotypes were counted. Pearson correlation coefficient was calculated between Hap1 and other haplotypes ($n \geq 40$) according to their sequence counts per day.

Identification of new canonical B.1.1.7 VOC genomic sequence

CGB is employed to identify a new canonical B.1.1.7 VOC genomic sequence(28) that carries the deletion g.a28271-. We first selected all the samples in the B.1.1.7 (CGB84017.91425) clade that carries all the B.1.1.7 characteristic mutations including g.a28271-. Then we filtered the samples by date and only kept the samples collected before 1 November, 2020. Viral strains with extra mutations were ignored. Then the sequence with accession EPI_ISL_629703, as the suggested new canonical B.1.1.7 VOC genomic sequence, is the first collected high-quality sequence without any extra mutations after the deletion g.a28271- occurred (Supplementary Figure 1).

Data and materials availability

All the SARS-CoV-2 data can be obtained from the Coronavirus GenBrowser and VENAS. Other data are available from the corresponding authors upon request.

Acknowledgments

We thank the researchers who generated and deposited sequence data of SARS-CoV-2 in GISAID, GenBank, CNGBdb, GWH, and NMDC making this study possible.

Funding

This work was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB38030100), the National Key Research and Development Project (Grant Nos. 2020YFC0847000, 2021YFC0863300, and 2020YFC0845900), and the National Natural Science Foundation of China (Grant No. 91531306). Financial support was also provided by the Shandong Academician Workstation Program #170401 (to G.P.Z.).

Author contributions

Conceptualization, J.Y., G.Z., D.Y., Y.H.P., B.S., Y.Z., G.P.Z., Y.L., H.L.; Data Analysis, J.Y., G.Z., D.Y., R.C., X.W., Y.L., C.Y., X.S.; Writing, J.Y., G.Z., D.Y., R.C., Y.L., Y.H.P., C.Y., X.S., Y.Z., G.P.Z., H.L.; Supervision & Funding Acquisition, G.Z., Y.Z., G.P.Z., Y.L., H.L.

Declaration of interests

The authors declare no competing interests.

Reference

1. Rambaut, A, Loman, N, Pybus, O, *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *virologicalorg*. 2020: 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.
2. Volz, E, Mishra, S, Chand, M, *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*. 2021.
3. Wu, F, Zhao, S, Yu, B, *et al.* A new coronavirus associated with human respiratory disease in China. *Nature*. 2020; **579**: 265-9.
4. Yu, D, Yang, X, Tang, B, *et al.* Coronavirus GenBrowser for monitoring the transmission and evolution of SARS-CoV-2. *medRxiv*. 2021: 2020.12.23.20248612.
5. Alpert, T, Brito, AF, Lasek-Nesselquist, E, *et al.* Early introductions and community transmission of SARS-CoV-2 variant B.1.1.7 in the United States. *medRxiv*. 2021.
6. Chen, C, Nadeau, S, Topolsky, I, *et al.* Quantification of the spread of SARS-CoV-2 variant B.1.1.7 in Switzerland. *medRxiv*. 2021.
7. Umair, M, Salman, M, Rehman, Z, *et al.* An upsurge of SARS CoV-2 B.1.1.7 variant in Pakistan. *medRxiv*. 2021.
8. Alcoba-Florez, J, Lorenzo-Salazar, JM, Gil-Campesino, H, *et al.* Monitoring the

rise of the SARS-CoV-2 lineage B.1.1.7 in Tenerife (Spain) since mid-December 2020. *medRxiv*. 2021.

9. Davies, NG, Jarvis, CI, Edmunds, WJ, *et al.* Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*. 2021.
10. Grint, DJ, Wing, K, Williamson, E, *et al.* Case fatality risk of the SARS-CoV-2 variant of concern B.1.1.7 in England. *medRxiv*. 2021.
11. Starr, TN, Greaney, AJ, Hilton, SK, *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*. 2020; **182**: 1295-310.
12. Tegally, H, Wilkinson, E, Giovanetti, M, *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*. 2021.
13. Korber, B, Fischer, WM, Gnanakaran, S, *et al.* Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020; **182**: 812-27.
14. Zhou, B, Thao, TTN, Hoffmann, D, *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature*. 2021.
15. V'kovski, P, Kratzel, A, Steiner, S, *et al.* Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol*. 2021; **19**: 155-70.
16. Lubinski, B, Tang, T, Daniel, S, *et al.* Functional evaluation of proteolytic activation for the SARS-CoV-2 variant B.1.1.7: role of the P681H mutation. *bioRxiv*. 2021.
17. Khan, A, Zia, T, Suleman, M, *et al.* Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: An insight from structural data. *Journal of cellular physiology*. 2021.
18. Liu, Y, Liu, J, Plante, KS, *et al.* The N501Y spike substitution enhances SARS-CoV-2 transmission. *bioRxiv*. 2021.
19. Hsiang, S, Allen, D, Annan-Phan, S, *et al.* The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*. 2020; **584**: 262-7.
20. Ling, Y, Cao, R, Qian, J, *et al.* An interactive viral genome evolution network analysis system enabling rapid large-scale molecular tracing of SARS-CoV-2. *bioRxiv*. 2020.
21. Tang, X, Wu, C, Li, X, *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev*. 2020; **7**: 1012-23.
22. Kozak, M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol*. 1987; **196**: 947-50.
23. Xu, K, Zheng, BJ, Zeng, R, *et al.* Severe acute respiratory syndrome coronavirus accessory protein 9b is a virion-associated protein. *Virology*. 2009; **388**: 279-85.
24. Ghorbani, A, Samarfard, S, Ramezani, A, *et al.* Quasi-species nature and differential gene expression of severe acute respiratory syndrome coronavirus 2 and phylogenetic analysis of a novel Iranian strain. *MEEGID*. 2020; **85**: 104556.
25. Bojkova, D, Klann, K, Koch, B, *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature*. 2020; **583**: 469-72.
26. Schelle, B, Karl, N, Ludewig, B, *et al.* Selective replication of coronavirus genomes that express nucleocapsid protein. *J Virol*. 2005; **79**: 6620-30.

27. Wu, J, Shi, YH, Pan, XY, *et al.* SARS-CoV-2 ORF9b inhibits RIG-I-MAVS antiviral signaling by interrupting K63-linked ubiquitination of NEMO. *Cell Rep*. 2021; **34**.
28. Chand, M, Hopkins, S, Dabrera, G, *et al.* Investigation of novel SARS-CoV-2 variant of concern 202012/01. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959438/Technical_Briefing_VOC_SH_NJL2_SH2.pdf. 2020.
29. Hadfield, J, Megill, C, Bell, SM, *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018; **34**: 4121-3.
30. Lam, HM, Ratmann, O, Boni, MF. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. *Mol Biol Evol*. 2018; **35**: 247-51.
31. Kim, Y, Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*. 2002; **160**: 765-77.
32. Li, H, Stephan, W. Maximum likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics*. 2005; **171**: 377-84.
33. Bouckaert, R, Vaughan, TG, Barido-Sottani, J, *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 2019; **15**(4).
34. Snijder, EJ, Limpens, RWAL, de Wilde, AH, *et al.* A unifying structural and functional model of the coronavirus replication organelle: Tracking down RNA synthesis. *Plos Biol*. 2020; **18**(6).

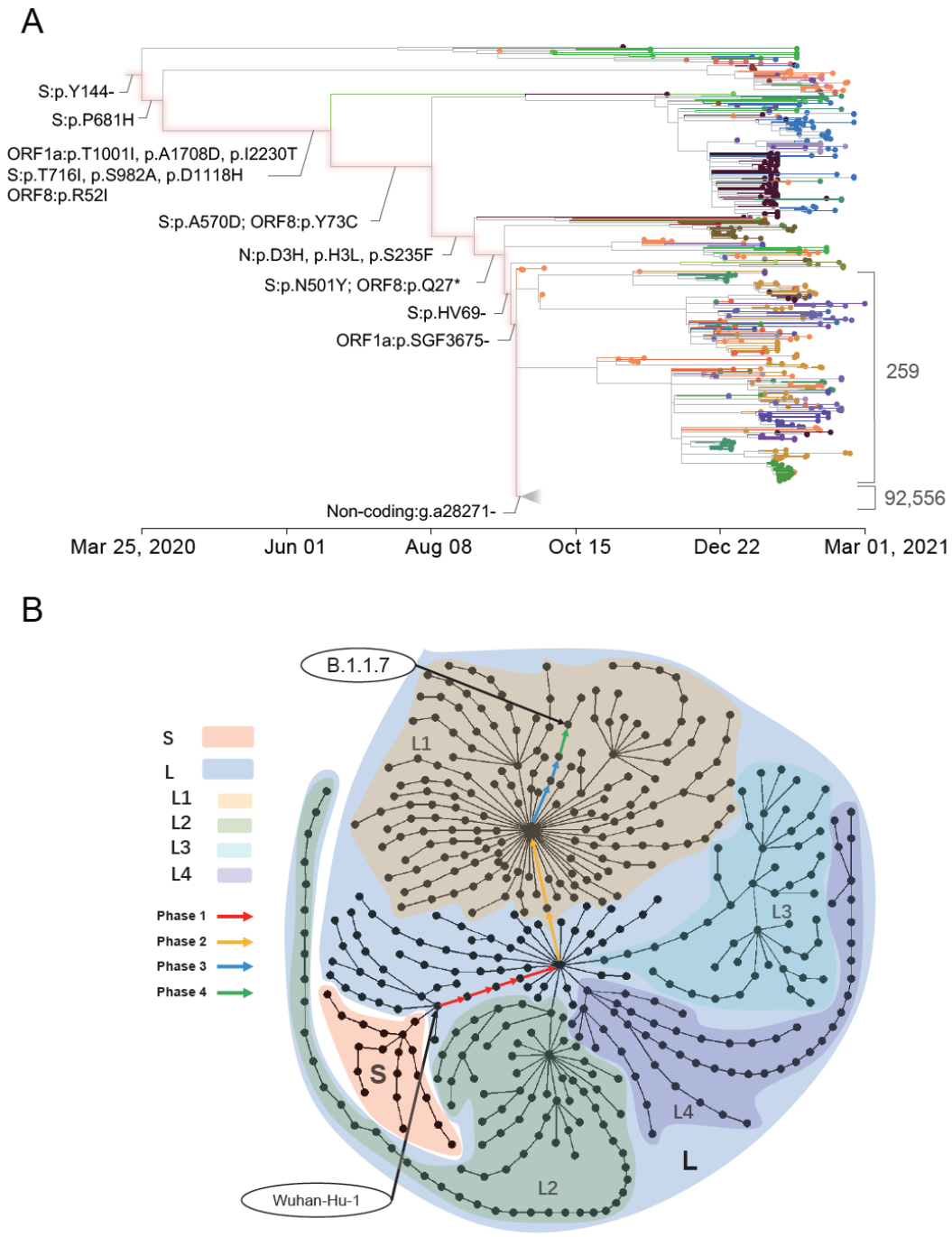


Figure 1. Sequential occurrence order of B.1.1.7 mutations.

(A) CGB evolutionary tree of SARS-CoV-2 lineage B.1.1.7. The B.1.1.7 lineage is highlighted in red and the major clade with 92,556 strains is collapsed. The analysis was performed on 400,051 high-quality SARS-CoV-2 genomic sequences using the Coronavirus GenBrowser (4). The CGB ID of the internal node with g.a28271- is CGB75056.269896. The number of B.1.1.7-like strains is 259, and the number of B.1.1.7 strains is 92,556.

(B) VENAS evolution network of SARS-CoV-2 by January 14, 2021. The dots represent the major genome types of SARS-CoV-2, and the lines between the dots are the evolutionary path formed by the combination of variants; the color shades represent the clades and subclades formed by genome types, where the L1

subclade is shaded in yellow; the L2 in green; the L3 in cyan, and the L4 in purple. The color arrows mark the evolutionary path from the most recent common ancestor of SARS-CoV2 to the B.1.1.7 linage, and four phases are indicated in different colors.

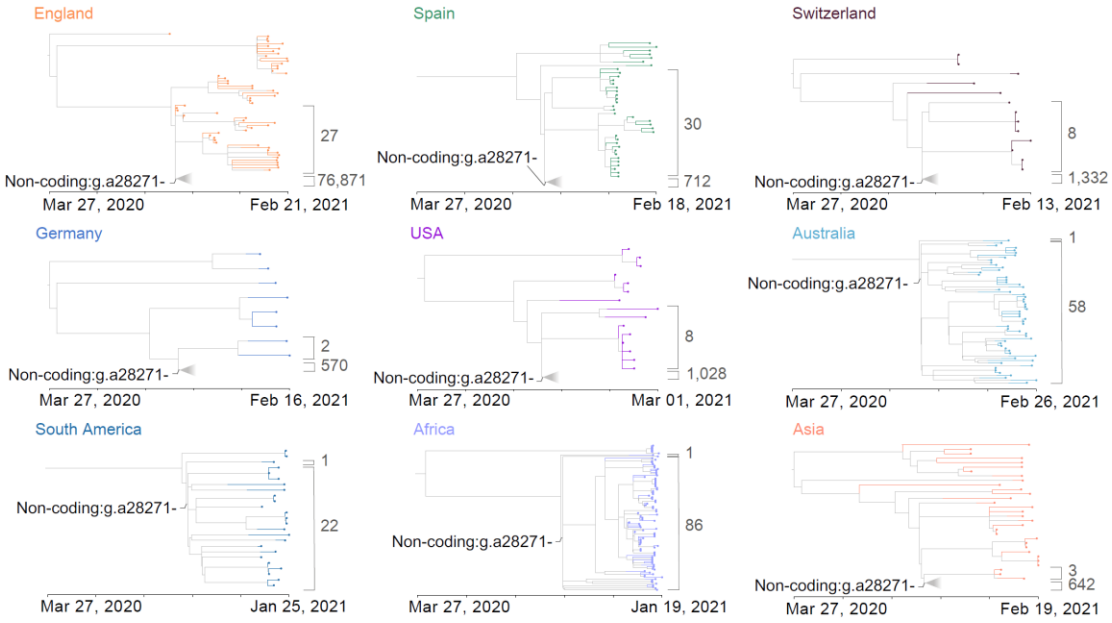


Figure 2. CGB evolutionary sub-trees of SARS-CoV-2 lineage B.1.1.7. Strains were collected from different countries/regions. The B.1.1.7 clade could be collapsed if its size is too large to be shown. For each sub-tree, the number of B.1.1.7-like (without the non-coding deletion) strains and that of B.1.1.7 (with the non-coding deletion) strains are labeled.

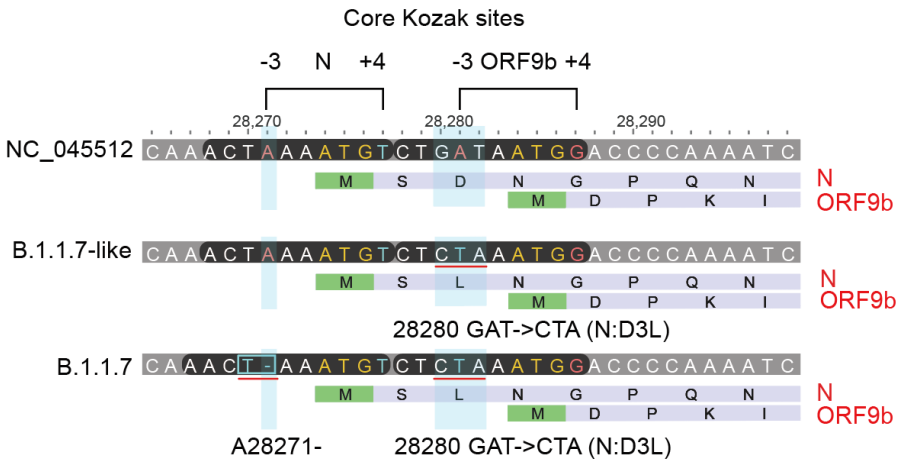


Figure 3. Two B.1.1.7 mutations change the core Kozak sites of *N* and *ORF9b* genes. The two positions -3 and +4 have the dominant influence (22). The grey bars are the nucleotide sequences of the variants. Two functional genes are presented under each sequence. Start codons are shown in green. The *N* and *ORF9b* genes with their amino acid sequences are colored in light purple. Sites that mutations happened are covered in light blue rectangle. The optimal Kozak sites are colored in red and non-optimal ones in light blue.

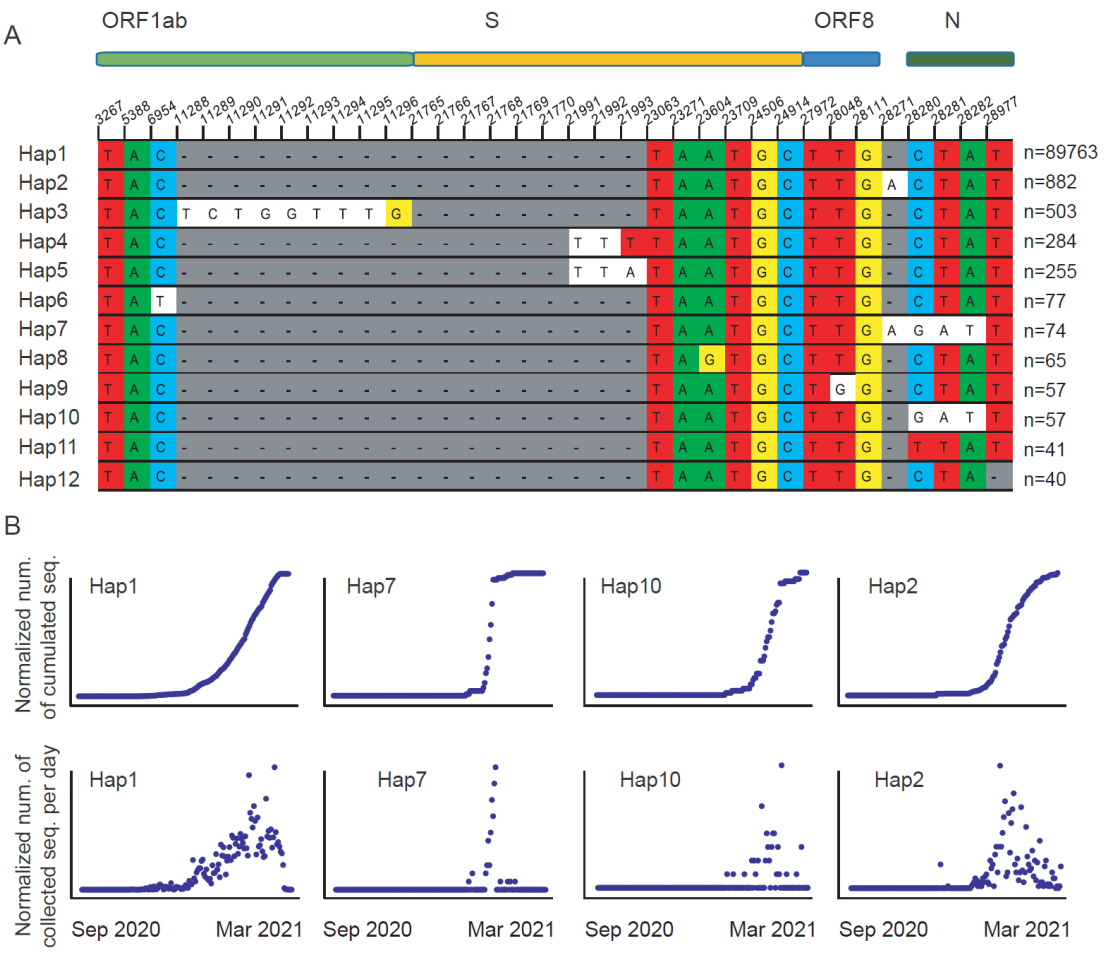


Figure 4. Haplotype information and its cumulative frequency within B.1.1.7 clade.

(A) Graphically display of haplotype details with site of considered mutations in ORF1ab, S, ORF8, a non-coding region, and N. Hap1 is the B.1.1.7 major haplotype. *n* is the number of strains for each haplotype, shown on the right of figure. Each colorful character is corresponding to a B.1.1.7 variant. Blank characters represent their ancestral states that are the same as those of the reference genome. Within Hap2 – 12, each haplotype contains at least one (back) mutation.

(B) The cumulative frequency of Hap1, Hap7, Hap10, and Hap2 in B.1.1.7 clade. The cumulative frequency of Hap7 shows the greatest deviation with that of Hap1. The y-axis is normalized among the four haplotypes.

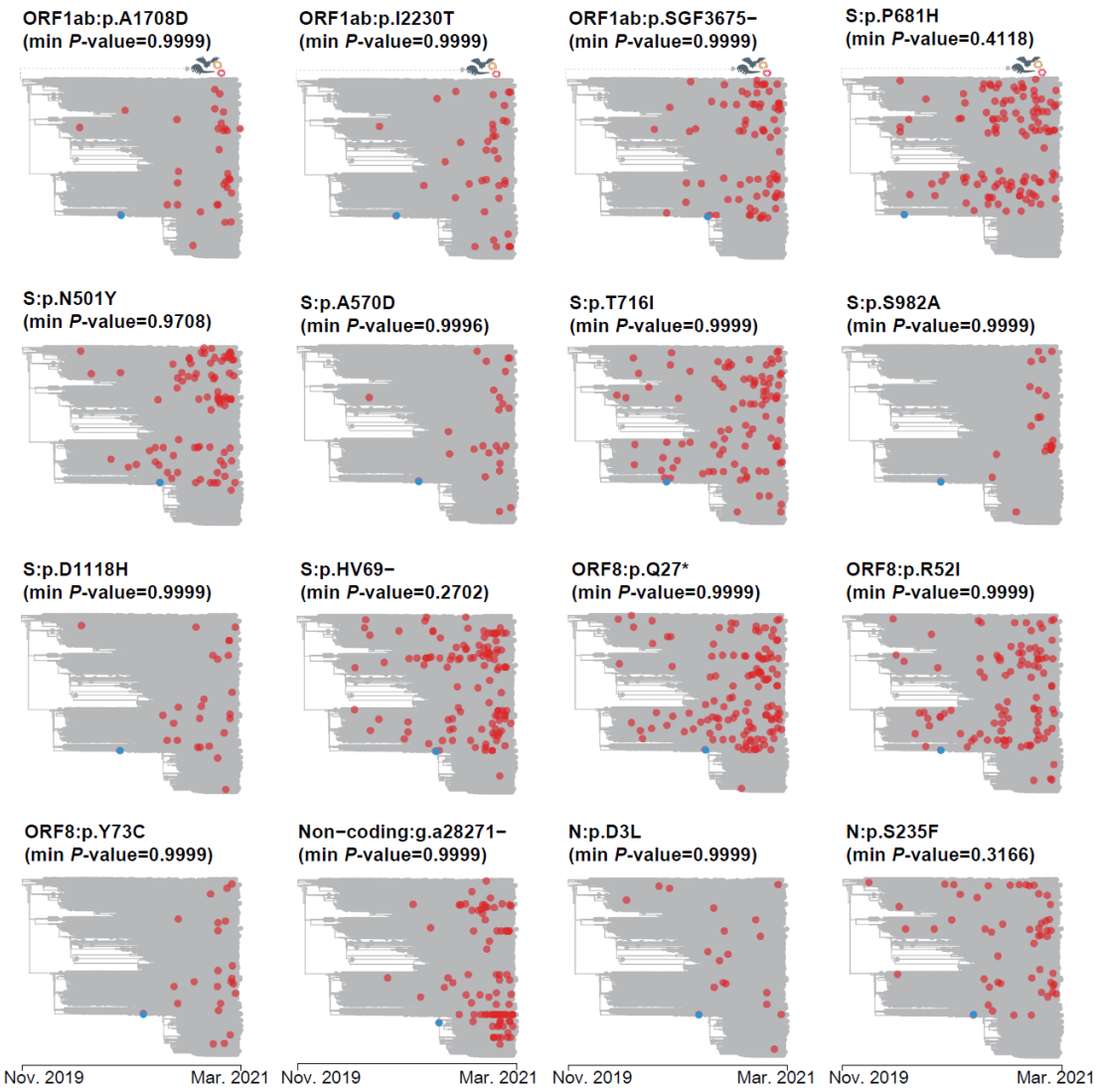


Figure 5. Recurrence of B.1.1.7 characteristic mutations in a huge evolutionary tree with 400,051 high-quality SARS-CoV-2 genomic sequences. The blue dots indicate mutations occurred on the B.1.1.7 lineage and the red dots indicate recurrent mutations. The min *P*-value marked is the minimal *P*-value testing whether the corresponding mutation alone is advantage in the spread of virus. All *P*-value showed are FDR corrected.

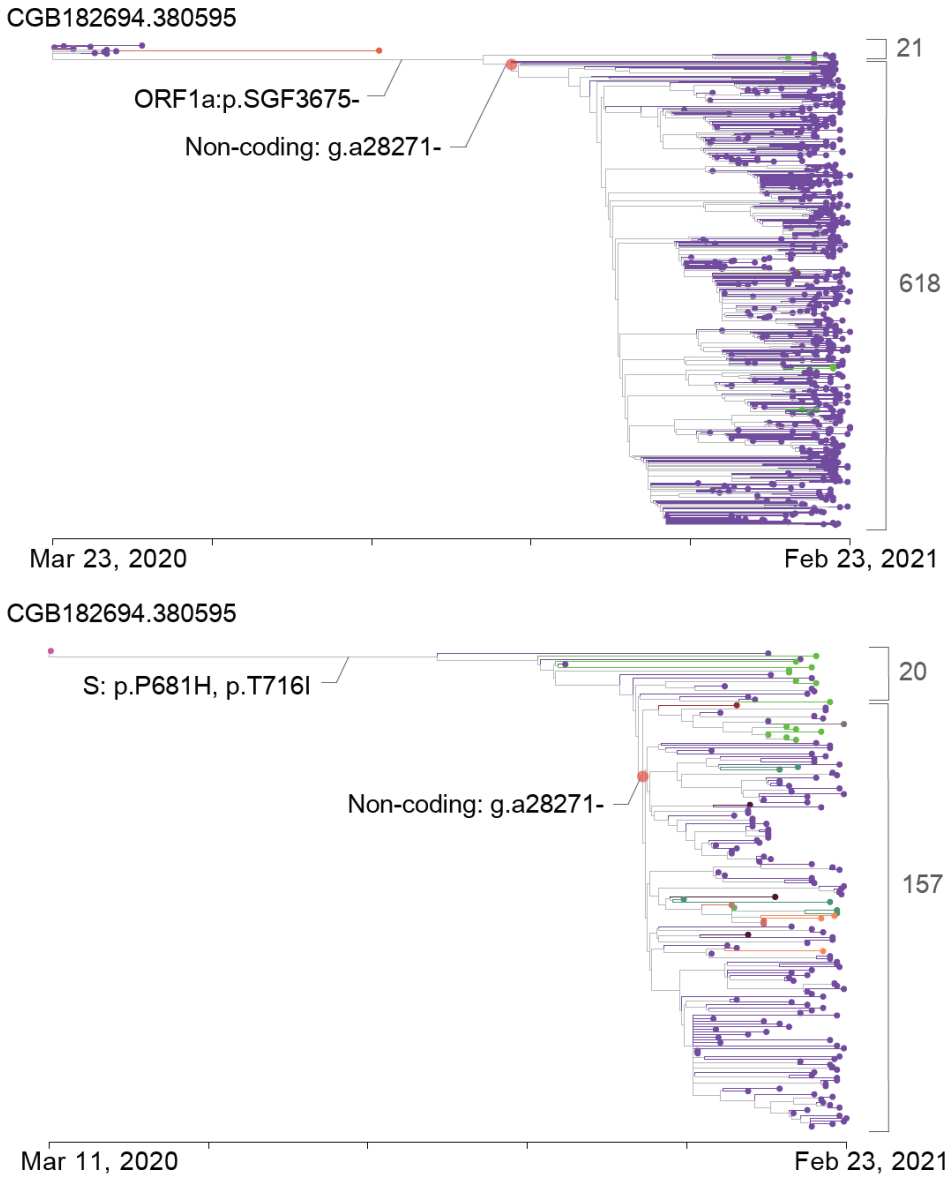
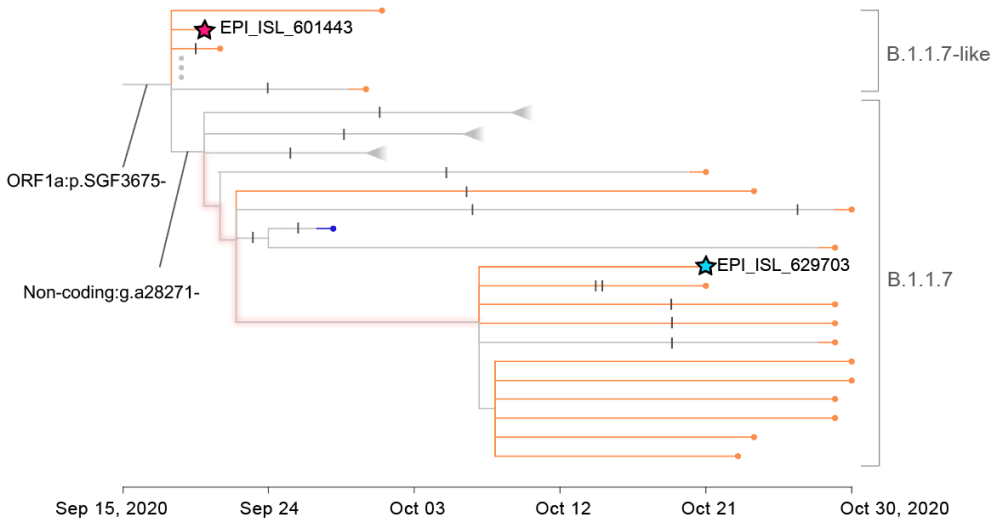


Figure 6. Two non-B.1.1.7 clades carrying the non-coding deletion g.a28271- and other B.1.1.7 characteristic mutations. All the internal nodes have been named by the CGB binary nomenclature system (4). The CGB ID of the expanding node (marked by red point) is presented on the top of tree. The B.1.1.7 mutations are marked.

Supplemental materials

Evolutionary insights into a non-coding deletion of SARS-CoV-2 B.1.1.7



Supplementary Figure 1. Suggested canonical B.1.1.7 VOC genomic sequence. The current canonical B.1.1.7 VOC genomic sequence is marked in red star. The new canonical B.1.1.7 VOC genomic sequence we suggested is in blue star. Each notch of the branches represents a mutation. The branches with no mutations are highlighted. The blue-star sequence is the first collected high-quality sequence without any extra mutations after the deletion g.a28271- occurred.

Supplementary Table 1. Characteristic mutations of B.1.1.7 lineage.

Development phase	Nucleotide mutation	Amino-acid mutation	Type
Phase4	g.c3267t	ORF1a:p.T1001I	SNV
	g.c5388a	ORF1a:p.A1708D	SNV
	g.t6954c	ORF1a:p.I2230T	SNV
	g. tctggtttt11288-	ORF1a:p.SGF3675-	INDEL
	g.tacatg21765-	S:p.HV69-	INDEL
	g.a23063t	S:p.N501Y	SNV
	g.c23271a	S:p.A570D	SNV
	g.c23604a	S:p.P681H	SNV
	g.c23709t	S:p.T716I	SNV
	g.t24506g	S:p.S982A	SNV
	g.g24914c	S:p.D1118H	SNV
	g.c27972t	ORF8:p.Q27*	SNV
	g.g28048t	ORF8:p.R52I	SNV
	g.a28111g	ORF8:p.Y73C	SNV
	g.a28271-	Non-coding	INDEL
	g.gat28280cta	N:p.D3L	SNV
	g.c28977t	N:p.S235F	SNV
Phase3	g.tta21991-	S:p.Y144-	INDEL
Phase2	g.ggg28881aac	N:p.RG203KR	SNV
Phase1	g.a23403g	S:p.D614G	SNV
	g.c14408t	ORF1a:p.P4715L	SNV

Supplementary Table 2. The correlation between the spread of B.1.1.7 major haplotype (Hap1) and that of mutated B.1.1.7 haplotypes.

Haplotype	The altered B.1.1.7 characteristic mutations	PCC¶
Hap1	None	1.0000
Hap7	g.a28271- and g.gat28280cta (N:p.D3L)	0.2456
Hap9	g.g28048t (ORF8:p.R52I)	0.2918
Hap12	g.c28977t (N:p.S235F)	0.4074
Hap10	g.gat28280cta (N:p.D3L)	0.4541
Hap11	g.gat28280cta (N:p.D3L)	0.5298
Hap8	g.c23604a (S: p.P681H)	0.5425
Hap2	g.a28271-	0.5739
Hap3	g.tctggtttt11288- (ORF1a:p.SGF3675-)	0.6484
Hap6	g.t6954c (ORF1a:p.I2230T)	0.6668
Hap4	g.tta21991- (S:p.Y144-)	0.7346
Hap5	g.tta21991- (S:p.Y144-)	0.7784

¶Pearson correlation coefficient is calculated between the sequence count per day of Hap1 and that of the *i*-th haplotype.