A Kozak-related non-coding deletion effectively increases B.1.1.7 transmissibility

Jianing Yang[1,†], Guoqing Zhang[1,†], Dalang Yu[1,†], Ruifang Cao[1,†], Xiaoxian Wu[2], Yunchao Ling[1], Yi-Hsuan Pan[4], Chunyan Yi[3], Xiaoyu Sun[3], Bing Sun[3], Yu Zhang[2], Guo-Ping Zhao[1,2,5,*], Yixue Li[1,6,*], Haipeng Li[1,7,8,*]

[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]Key Laboratory of Brain Functional Genomics of Ministry of Education, School of Life Science, East China Normal University, Shanghai 200062, 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]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China.

[8]Lead Contact

[†]These authors contributed equally.

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

Abstract

The high transmissibility acquisition of SARS-CoV-2 Variant of Concern (VOC) B.1.1.7 remains unclear and only mutations in coding regions have been examined. We analyzed 875,338 high-quality SARS-CoV-2 genomic sequences and the epidemiology metadata. The occurrence of a non-coding deletion (g.a28271-) in the B.1.1.7 background immediately causes the rapid spread of B.1.1.7. The number of B.1.1.7-like strains lacking the deletion is significantly less than that of B.1.1.7 strains ($n = 259$ *vs* 92,688, *P*-value$< 4.9 \times 10^{-324}$). The same highly significant statistics is observed in different countries, gender and age groups. However, the deletion alone does not cause such high viral transmissibility. The deletion and another mutation (g.gat28280cta) co-affect translational efficiency of the genes *N* and *ORF9b* by changing the core Kozak sites. The deletion interacts synergistically with S:p.P681H and S:p.T716I to increase viral transmissibility. Therefore, the Kozak-related non-coding deletion, also carried by the Delta VOC, is crucial for the high viral transmissibility of SARS-CoV-2.

Introduction

SARS-CoV-2 lineage B.1.1.7, also known as Variant of Concern (VOC) 202012/01 or the Alpha VOC, is a variant first detected in the UK in September 2020 (*1*) and has higher transmissibility than the preexisting variants (*2*). Its high transmissibility remains similar across different age, sex and socioeconomic strata (*3*). It has spread to 97 countries/regions within seven months and its global infection frequency increases quickly to over 80% (Supplementary Figure S1). Comparing with the reference genomic sequence of SARS-CoV-2 (GenBank accession number: NC_045512.2) (*4*), the sequence of the B.1.1.7 variant has 20 non-synonymous mutations and amino-acid deletions in *ORF1ab*, spike (*S*), *ORF8*, and nucleocapsid (*N*) genes (Supplementary Table S1) (*5*). Among them, it was previously found that each of the mutations S:p.N501Y (*6, 7*) and S:p.D614G (*8, 9*) may increase the viral transmissibility, and S:p.P681H, located on the spike S1/S2 cleavage site, may affect the cleavableness and activation of the spike protein (*1, 10*). However, the crucial mutations for the high transmissibility of the B.1.1.7 VOC still remain unclear.

Most if not all of the current studies focused merely on non-synonymous mutations and amino-acid deletions(*11-16*) when studying the crucial mutations for the high transmissibility of the B.1.1.7 VOC. Non-coding mutations have been ignored in those studies and are not presented in the pathogen genomics platform Nextstrain (www.nextstrain.org) either (*17*). Therefore, we analyzed 875,338 high-quality SARS-CoV-2 genomic sequences and the associated epidemiology metadata. The occurrence of a non-coding deletion (g.a28271-) in the B.1.1.7 background immediately causes the rapid spread of B.1.1.7 VOC. Although the B.1.1.7 spike appears to have a higher binding affinity with the angiotensin-converting enzyme 2 (ACE2) (*11*), the variant with B.1.1.7 spike had no high transmissibility until the non-coding deletion occurred. Interestingly, the non-coding deletion alone does not cause such high viral transmissibility and is unlikely to have apparent fitness advantage, indicating the importance of mutation interactions. We also found that the non-coding deletion is carried by the Delta VOC.

Therefore, the Kozak-related non-coding deletion and its interactions with other mutations are crucial for the high viral transmissibility of SARS-CoV-2.

Methods

Data sources

The annotated evolutionary tree and evolutionary network data were obtained from the Coronavirus GenBrowser (*18*) and VENAS (*19*). All sequence data of SARS-CoV-2 were obtained from the 2019nCoVR database (*20*, *21*), which is an integrated resource based on Global Initiative on Sharing All Influenza Data (GISAID) (*22*, *23*), National Center for Biotechnology Information (NCBI) GenBank (*24*), China National GeneBank DataBase (CNGBdb) (*25*), the Genome Warehouse (GWH) (*26*), and the National Microbiology Data Center (NMDC, https://nmdc.cn/).

The effectiveness of the non-coding deletion in improving transmissibility

To evaluate the effectiveness of the non-coding deletion in improving transmissibility, we tested whether B.1.1.7-like and B.1.1.7 strains have the same transmissibility. The former indicate the viral strains lacking the non-coding deletion but carrying all other characteristic mutations of B.1.1.7 (Figure 1A, Supplementary Table S1) (*5*), including all B.1.1.7 spike mutations(*11*). The latter carry all those mutations, including the non-coding deletion. The null hypothesis is that the B.1.1.7-like and B.1.1.7 strains have the same transmissibility, and the alternative hypothesis is that the B.1.1.7-like strains have lower transmissibility than the B.1.1.7 strains. The binomial probability was used to test the null hypothesis, and the test was one-tailed. The analysis was based on the data version “data.2021-03-06” ($n = 400{,}051$) of Coronavirus GenBrowser (CGB) (*18*), where *n* is the number of viral strains.

The difference between the first appearance time of the B.1.1.7-like and that of the B.1.1.7 strain is small (Figures 1A, 2A). Therefore, we set the probability to observe a B.1.1.7-like or a B.1.1.7 strain as 0.5 under the null hypothesis. This is a

conservative treatment since the B.1.1.7-like strain emerged before the B.1.1.7 strain, thus the former had more time to spread than the latter.

Reappearance of mutations in the evolutionary tree

To examine the reappearance of mutations, recurrent mutations and mutations due to recombination were considered. Considering the degeneracy of the genetic code, we searched amino acid mutations by using the form of amino acid change, instead of that of nucleotide change in the CGB (*18*). To search the non-coding deletion (g.a28271-) in the evolutionary tree, we used the string "A28271-". To present the reappearance patterns of mutations, the data version "data.2021-03-06" ($n = 400{,}051$) of the CGB (*18*) was used. This data was also used to examine the frequency trajectory of a B.1.1.7 characteristic mutation after the B.1.1.7 strains were excluded.

Identification of new canonical B.1.1.7 genomic sequence

The CGB was employed to identify a new canonical B.1.1.7 genomic sequence (*5*) that carries the deletion g.a28271- and all other B.1.1.7 characteristic mutations (Figure 1A, Supplementary Table S1). We first selected all the strains 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 strains by date and only kept the strains 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 genomic sequence, is the first collected high-quality sequence without any extra mutations after the deletion g.a28271- occurred (Supplementary Figure S2).

Results

A crucial non-coding deletion in the B.1.1.7 lineage

The sequential occurrence order of B.1.1.7 characteristic mutations may provide the important clues to identify the crucial mutations for the B.1.1.7 high

transmissibility. Therefore, the B.1.1.7 lineage was examined using the Coronavirus GenBrowser (CGB) (*18*). The CGB evolutionary tree shows the sequential occurrence of B.1.1.7 characteristic mutations (Figure 1A). Interestingly, the results indicate that, after all the B.1.1.7 characteristic amino acid mutations occurred, the rapid spread of virus was not observed until a non-coding deletion occurred. To confirm the evolutionary path of B.1.1.7, we applied VENAS (*19*) to obtain an evolution network of SARS-CoV-2 major haplotypes (Figure 1B). The results are consistent with that of CGB evolutionary tree based analysis. Therefore, the occurrence of a non-coding deletion (g.a28271-), accompanied with other B.1.1.7 amino acid changes, immediately causes the rapid spread of B.1.1.7 VOC.

The non-coding deletion effectively increases the transmissibility of B.1.1.7

The occurrence of the non-coding deletion, located between *ORF8* and *N* genes, immediately causes the rapid spread of B.1.1.7 VOC (Figure 1A). To evaluate the effectiveness of the non-coding deletion in increasing the viral transmissibility, we compared the number of B.1.1.7-like strains, *i.e.*, lacking the non-coding deletion but carrying all other characteristic mutations of B.1.1.7 (*5*), with that of B.1.1.7 strains. Their numbers are highly significantly different ($n = 259$ *vs* 92,688, *P*-value$< 4.9 \times 10^{-324}$), indicating that B.1.1.7-like strains do not demonstrate a high transmissibility as B.1.1.7 strains do. Therefore, the non-coding deletion g.a28271- may contribute markedly to increase the transmissibility of B.1.1.7.

Pooling data of viral sequences from different countries is likely to be biased due to complex differences in sampling with respect to either viral genome sequencing capacities or anti-contagion policies on the pandemic among the targeted countries (*27*). To address this problem, the numbers of B.1.1.7-like and B.1.1.7 strains were pairwise compared for individual countries and continents (Figure 2A), i.e., England (27 *vs* 76,871), Spain (30 *vs* 712), Switzerland (8 *vs* 1,332), Germany (2 *vs* 570), USA (8 *vs* 1,028), Australia (1 *vs* 58), South America (1 *vs* 22), Africa (1 *vs* 86), and Asia (3 *vs* 642). The transmissibility of strains without or with the non-coding deletion is

significantly unequal (Table 1, *P*-value$\leq 2.74 \times 10^{-6}$). The same highly significant statistics was observed in 10 more countries, such as India and Italy. Moreover, the same conclusion holds when considering different gender and age groups (Supplementary Tables S2, S3). Therefore, the non-coding deletion g.a28271- effectively increases the transmissibility of B.1.1.7.

The g.a28271- and g.gat28280cta change the core Kozak sites of *N* and *ORF9b* genes

The base 28,271 is located at the third base upstream of the start codon of the *N* gene, whose expression is associated with the viral replication and has the highest translational rate (*28, 29*). The g.a28271- 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 2B) (*30*). When the homological site of the SARS-CoV genome was mutated to another undesirable one (C at -3, T at +4), the expression of N protein was reduced and the translation of ORF9b protein increased (*31*). The ORF9b protein was found to be translated via a leaky ribosomal scanning mechanism (*31*), and has an interferon (IFN) antagonistic activity and can suppress the IFN production (*32*). A recent proteomics survey found that the B.1.1.7 VOC has dramatically increased protein level of ORF9b (*12*), which is consistent with the function of the Kozak-related non-coding deletion.

Another B.1.1.7 mutation g.gat28280cta (N:p.D3L) at the ninth base downstream of the deletion g.a28271- changes the Kozak core sequence of *ORF9b* (Figure 2B) (*30*). It is expected that the expression level of *ORF9b* protein may be affected (*30*). However, this remains to be determined because of the leaky ribosomal scanning effect (*31*). Overall, these two mutations change the core Kozak sites and may co-affect the translational efficiency of gene *N* and *ORF9b*.

High viral transmissibility associated with multiple B.1.1.7 mutations

Besides the non-coding deletion, there are 16 non-synonymous mutations and

amino-acid deletions occurred recently along the B.1.1.7 lineage (Figure 1A), including S:p.N501Y and S:p.P681H. We then examined whether each of those mutations alone could increase the viral transmissibility in the background of the D614G substitution. Since all these mutations have appeared multiple times in the genome of SARS-CoV-2 (Supplementary Figure S3), we checked the frequency trajectory of each mutation when the B.1.1.7 lineage was excluded. We did not find a rapid frequency growth (Supplementary Figure S4), indicating that each of these 17 mutations alone is not associated with high viral transmissibility since the pandemic. Thus it is very unlikely that the high transmissibility of B.1.1.7 is caused by a single mutation.

We then searched the variants carrying the non-coding deletion and other 16 B.1.1.7 characteristic mutations (Figure 1A) in non-B.1.1.7 clades. Clades were chosen only if the occurrence of a B.1.1.7 characteristic mutation immediately leads a relatively rapid spread of virus. The largest clade (CGB199165.262639) is evidential to the synergistical effect among its associated mutations (Figure 3) in the background of the D614G substitution (*8, 9*). The two mutations (S:p.P681H, and S:p.T716I) first occur, and no rapid spread is observed until g.a28271- occurs. The variant with the first two mutations appears to spread significantly slower than the triple-mutated variant ($n = 59$ *vs* 1,196, *P*-value $= 1.92 \times 10^{-276}$). The conclusion remains the same when only considering the strains collected from the USA ($n = 43$ *vs* 1,118, *P*-value$= 1.47 \times 10^{-271}$). This observation suggests that g.a28271- may interact synergistically with one, or both of S:p.P681H and S:p.T716I to increase the viral transmissibility.

Discussion

In this study, we find that the non-coding deletion g.a28271- plays an essential role in the high transmissibility of B.1.1.7 VOC. It has been documented that the B.1.1.7 spike improves the angiotensin-converting enzyme 2 (ACE2) affinity for about 5-fold, comparing with the D614G spike (*11*). However, the epidemiological

data show that this increase of ACE2 affinity cannot cause the high transmissibility of B.1.1.7 VOC when the non-coding deletion g.a28271- is lack.

Sequence with accession EPI_ISL_601443 was previously recommended to be the canonical B.1.1.7 genomic sequence (*5*). However, it does not carry the crucial non-coding deletion g.a28271-. The deletion is not presented in the pathogen genomics platform Nextstrain either (www.nextstrain.org) (*17*) because it is non-coding. Therefore, to investigate the viral transmissibility, we suggest using the sequence with accession EPI_ISL_629703 as the canonical B.1.1.7 genomic sequence (collected 21 October, 2020, in the UK) (Supplementary Figure S2).

Interestingly, it is likely that the deletion g.a28271- occurs due to recurrent mutation instead of recombination in the B.1.1.7 lineage. First, the probability of occurring g.a28271- is high. There are four continuous 'A' nucleotides between 28,271 and 28,274 (Figure 2B). When one of these nucleotides is deleted, it causes the same effect. All of those deletions are categorized as g.a28271- in the CGB. Thus, the deletion rate is roughly quadrupled. Second, there is only one mutation g.a28271- on the identified branch (CGB84017.91425). Recombination tends to create a hybrid genomic structure (*18, 33*). The two previously mutated alleles (g28111, cta28280) remain unchanged when the mutation g.a28271- occurs although both mutated alleles are next to the genomic position 28,271. Therefore, g.a28271- may be occurred as recurrent mutation in the B.1.1.7 clade.

Genomic mutations related to the transmissibility of a pandemic etiological pathogen such as SARS-CoV-2 is complex and difficult to be revealed merely *via* genetic analysis with limited and incomplete supporting data of epidemiology. However, this study unveils a few of the crucial mutations, *S*-gene and other genes, non-synonymous and non-coding mutations of B.1.1.7, all likely affect the transmissibility synergistically as a beneficial haplotype. Moreover, g.a28271- was also found in the Delta VOC, known as the Indian VOC or B.1.617.2 (Supplementary Figure S5). Overall, our analyses indicate that non-coding mutations can be crucial for viral transmissibility by altering translational efficiency and interacting with other

mutations.

Acknowledgments

We thank the researchers who generated and deposited sequence data of SARS-CoV-2 in GISAID, GenBank, CNGBdb, GWH, and NMDC. 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), the National Natural Science Foundation of China (Grant No. 91531306), and the Shandong Academician Workstation Program #170401 (to G.P.Z.).

References

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; 593: 266-9.
3. Davies, NG, Abbott, S, Barnard, RC, *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science. 2021; 372: eabg3055.
4. Wu, F, Zhao, S, Yu, B, *et al.* A new coronavirus associated with human respiratory disease in China. Nature. 2020; 579: 265-9.
5. 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.
6. 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.
7. Tegally, H, Wilkinson, E, Giovanetti, M, *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 2021; 592: 438-43.
8. Zhou, B, Thao, TTN, Hoffmann, D, *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. Nature. 2021; 592: 122-7.
9. 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.
10. 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.
11. Gobeil, SM-C, Janowska, K, McDowell, S, *et al.* Effect of natural mutations of

SARS-CoV-2 on spike structure, conformation, and antigenicity. Science. 2021.

12. Thorne, LG, Bouhaddou, M, Reuschl, A-K, *et al.* Evolution of enhanced innate immune evasion by the SARS-CoV-2 B.1.1.7 UK variant. bioRxiv. 2021.
13. 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.
14. 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.
15. Liu, Y, Liu, J, Plante, KS, *et al.* The N501Y spike substitution enhances SARS-CoV-2 transmission. bioRxiv. 2021.
16. Cai, Y, Zhang, J, Xiao, T, *et al.* Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. Science. 2021: eabi9745.
17. Hadfield, J, Megill, C, Bell, SM, *et al.* Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018; 34: 4121-3.
18. Yu, D, Yang, X, Tang, B, *et al.* Coronavirus GenBrowser for monitoring the transmission and evolution of SARS-CoV-2. medRxiv. 2021.
19. 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.
20. Zhao, W-M, Song, S-H, Chen, M-L, *et al.* The 2019 novel coronavirus resource. Hereditas (Beijing). 2020; 42(2): 212-21.
21. Gong, Z, Zhu, J-W, Li, C-P, *et al.* An online coronavirus analysis platform from the National Genomics Data Center. Zool Res. 2020; 41(6): 705-8.
22. Elbe, S, Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall. 2017; 1(1): 33-46.
23. Shu, YL, McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Eurosurveillance. 2017; 22(13): 2-4.
24. Sayers, EW, Beck, J, Bolton, EE, *et al.* Database resources of the National Center

for Biotechnology Information. Nucleic Acids Res. 2021; 49: D10-D7.

25. Chen, F, You, L, Yang, F, *et al.* CNGBdb: China National GeneBank DataBase. Hereditas (Beijing). 2020; 42(8): 799-809.
26. Zhang, Z, Zhao, W, Xiao, J, *et al.* Database resources of the National Genomics Data Center in 2020. Nucleic Acids Res. 2020; 48(D1): D24-D33.
27. 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.
28. Bojkova, D, Klann, K, Koch, B, *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. Nature. 2020; 583: 469-72.
29. Schelle, B, Karl, N, Ludewig, B, *et al.* Selective replication of coronavirus genomes that express nucleocapsid protein. J Virol. 2005; 79: 6620-30.
30. Kozak, M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Biol. 1987; 196: 947-50.
31. 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.
32. 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: 108761.
33. Lam, HM, Ratmann, O, Boni, MF. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. Mol Biol Evol. 2018; 35: 247-51.
34. 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.

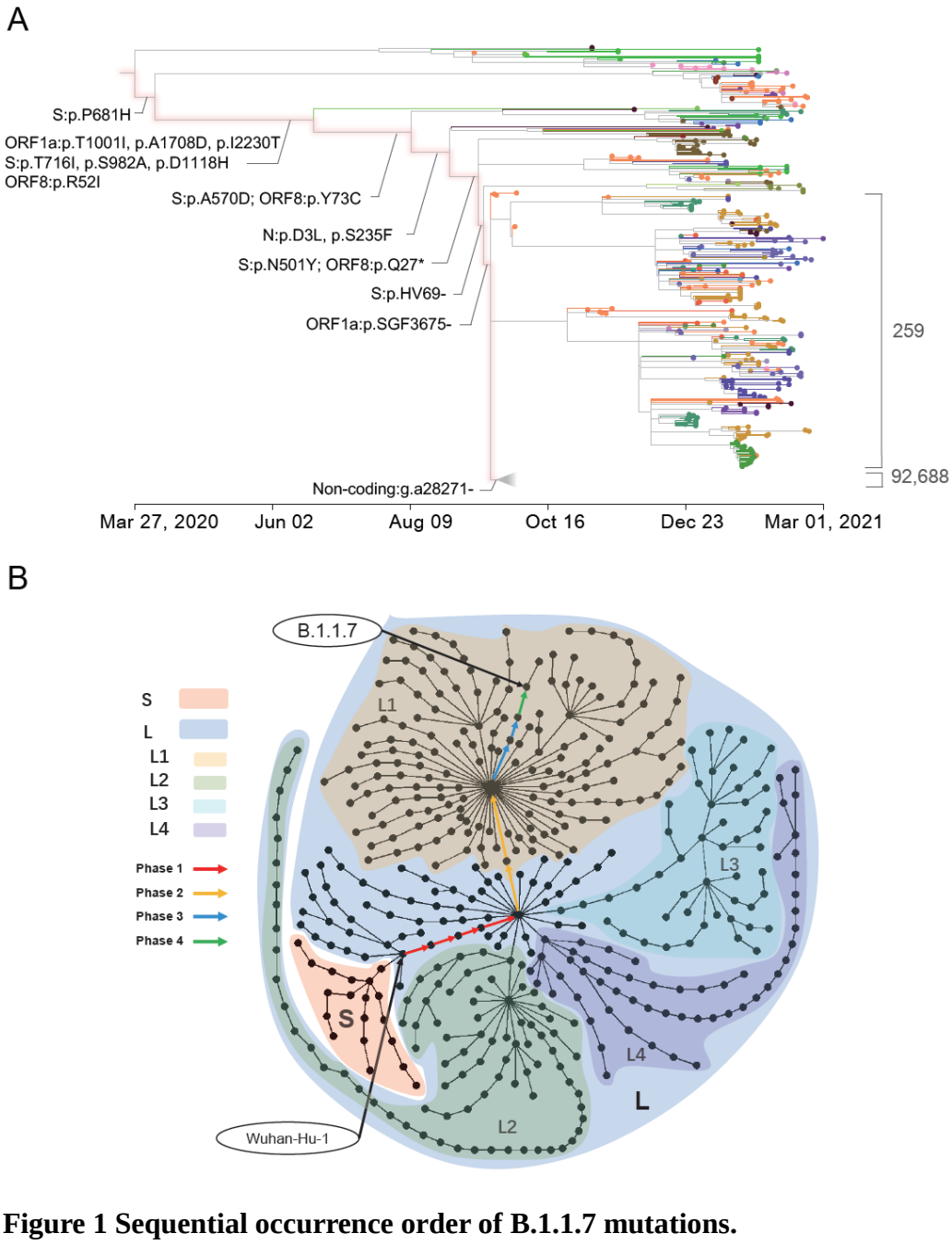

Figure 1 Sequential occurrence order of B.1.1.7 mutations.

(A) CGB evolutionary tree of SARS-CoV-2 lineage B.1.1.7. The analysis was performed on 400,051 high-quality SARS-CoV-2 genomic sequences using the Coronavirus GenBrowser (*18*). The searchable CGB ID of the internal node with g.a28271- is CGB84017.91425, assigned by the CGB binary nomenclature system. The B.1.1.7 clade was collapsed. The mutations on the highlighted branches were labeled. The number of B.1.1.7-like strains is 259, and the number of B.1.1.7 strains is 92,688.

(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 L/S naming system follows the previous study (*34*). 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.

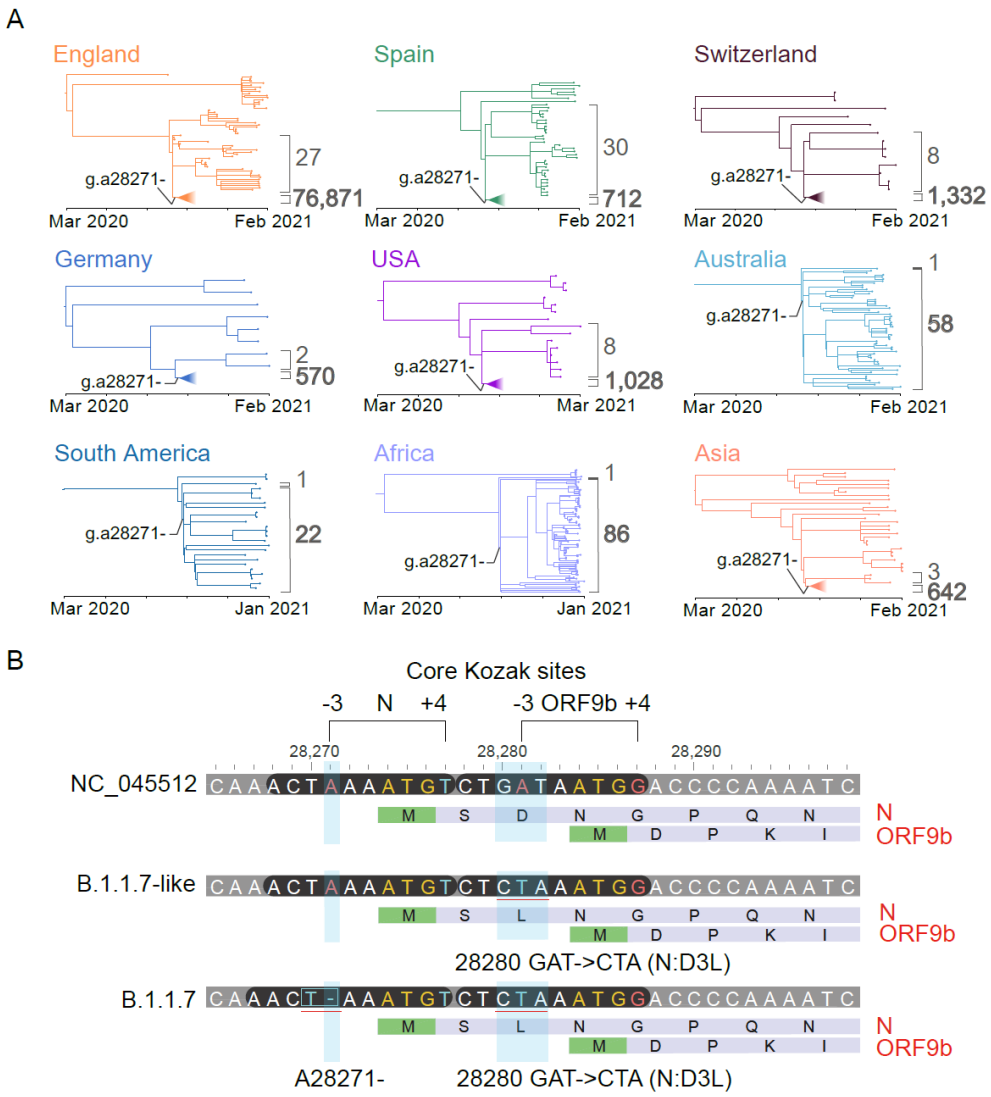

Figure 2. A Kozak-related non-coding deletion g.a28271- is essential for the high transmissibility of the B.1.1.7 VOC.

(A) The different transmissibility between the B.1.1.7-like and the B.1.1.7 strains *via* employing the CGB (*18*). Strains were filtered for different countries or continents. The B.1.1.7 clade was collapsed if its size was too large to be shown. For each sub-tree, the plain number of B.1.1.7-like (without the non-coding deletion) strains and the bold number of B.1.1.7 (with the deletion) strains are labeled.

(B) 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 (*30*). 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

371 their amino acid sequences are colored in light purple. Sites that mutations
372 happened are covered in light blue rectangle. The optimal Kozak sites are colored
373 in red and non-optimal ones in light blue.

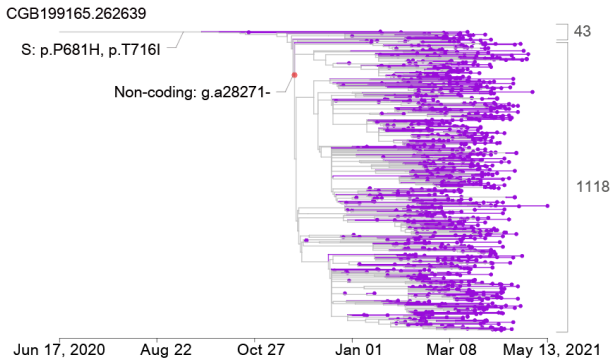

Figure 3. A non-B.1.1.7 rapid expanding clade carrying the non-coding deletion g.a28271- and two B.1.1.7 characteristic mutations in the background of the D614G substitution.

The searchable CGB ID of the expanding node (marked by red point) was presented on the top of tree. The B.1.1.7 mutations were marked. The data version "data.2021-05-20" ($n = 875{,}338$) of the CGB (*18*) was used. The strains collected from the USA were shown.

383 **Table 1. The number of B.1.1.7-like and B.1.1.7 strains in different countries and**
384 **continents.**

Country/continent*	The number of strains		*P*-value
	B.1.1.7-like	B.1.1.7	
England†	27	76,871	$< 4.9 \times 10^{-324}$
Spain †	30	712	1.16×10^{-170}
Switzerland†	8	1,332	$< 4.9 \times 10^{-324}$
Germany †	2	570	1.06×10^{-167}
USA †	8	1,028	4.35×10^{-293}
Australia †	1	58	1.02×10^{-16}
Norway	1	210	6.41×10^{-62}
Denmark	1	4,494	$< 4.9 \times 10^{-324}$
India	2	16	5.84×10^{-4}
Ireland	2	897	9.55×10^{-266}
France	2	1,059	2.27×10^{-314}
Sweden	16	182	3.56×10^{-37}
Finland	26	198	2.60×10^{-34}
Austria	28	242	4.85×10^{-44}
Italy	29	734	5.33×10^{-178}
Belgium	72	1,230	4.52×10^{-273}
South America †	1	22	2.74×10^{-6}
Africa †	1	86	5.62×10^{-25}
Asia †	3	642	3.05×10^{-187}

385 *Countries/continents with more than 10 viral strains (B.1.1.7-like and B.1.1.7).

386 †These six countries and three continents were shown in Figure 2A.

387