Identification of common deletions in the spike protein of SARS-CoV-2

Zhe Liu[*1,2], Huanying Zheng[*2], Huifang Lin[*1,2], Mingyue Li[3], Runyu Yuan[1,2], Jingju Peng[1,2], Qianlin Xiong[1,2], Jiufeng Sun[1,2], Baisheng Li[2], Jie Wu[2], Lina Yi[1,2], Xiaofang Peng[1,2], Huan Zhang[1,2], Wei Zhang[1,2], Ruben J.G. Hulswit[4], Nick Loman[5], Andrew Rambaut[6], Changwen Ke[2], Thomas A. Bowden[4], Oliver G Pybus[7], Jing Lu[1,2]

Affiliations:

[1] Guangdong Provincial Institution of Public Health, Guangzhou, China;

[2] Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China;

[3]Department of Rehabilitation Medicine, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

[4]Division of Structural Biology, Welcome Centre for Human Genetics, University of Oxford, Oxford, UK

[5]Institute of Microbiology and Infection, University of Birmingham, UK

[6]Institute of Evolutionary Biology, University of Edinburgh, UK

[7]Department of Zoology, University of Oxford, Oxford, UK

[*] Zhe Liu, Huanying Zheng, Huifang Lin contributed equally to this work. Author order was determined in order of increasing seniority.

Correspondence to: Oliver G Pybus, oliver.pybus@zoo.ox.ac.uk or Jing Lu, Jimlu0331@163.com.

Abstract

SARS-CoV-2 is a novel coronavirus first identified in December 2019. Notable features make SARS-CoV-2 distinct from most other previously-identified Betacoronaviruses, including the receptor binding domain of SARS-CoV-2 and a unique insertion of twelve nucleotide or four amino acids (PRRA) at the S1/S2 boundary. In this study, we identified two deletion variants of SARS-CoV-2 that either directly affect the polybasic cleavage site itself (NSPRRAR) or a flanking sequence (QTQTN). These deletions were verified by multiple sequencing methods. *In vitro* results showed that the deletion of NSPRRAR likely does not affect virus replication in Vero and Vero-E6 cells, however the deletion of QTQTN may restrict late phase viral replication. The deletion of QTQTN was detected in 3 of 68 clinical samples and half of 24 *in vitro* isolated viruses, whilst the deletion of NSPRRAR was identified in 3 *in vitro* isolated viruses. Our data indicate that (i) there may be distinct selection pressures on SARS-CoV-2 replication or infection *in vitro* and *in vivo*, (ii) an efficient mechanism for deleting this region from the viral genome may exist, given that the deletion variant is commonly detected after two rounds of cell passage, and (iii) the PRRA insertion, which is unique to SARS-CoV-2, is not fixed during virus replication *in vitro*. These findings provide information to aid further investigation of SARS-CoV-2 infection mechanisms and a better understanding of the NSPRRAR deletion variant observed here.

Important notes

The spike protein determines the infectivity and host range of coronaviruses. SARS-CoV-2 has two unique features in its spike protein, the receptor binding domain and an insertion of twelve nucleotides at the S1/S2 boundary resulting a furin-like cleavage site. Here, we identified two deletion variants of

SARS-CoV-2 that either directly affect the furin-like cleavage site itself (NSPRRAR) or a flanking sequence (QTQTN) and investigated these deletions in cell isolates and clinical samples. The absence of the polybasic cleavage site in SARS-CoV-2 did not affect virus replication in Vero or Vero-E6 cells. Our data indicate the PRRAR and its flanking sites are not fixed *in vitro*, thus there appears to be distinct selection pressures on SARS-CoV-2 sequences *in vitro* and *in vivo*. Further investigation of the mechanism of generating these deletion variants and their infectivity in different animal models would improve our understanding of the origin and evolution of this virus.

Introduction

SARS-CoV-2 is a novel coronavirus that was first identified at the end of December 2019 (1) and responsible for the global pandemic of COVID-19(2). Unlike the two other zoonotic coronaviruses, SARS-CoV-1 and MERS-CoV(3), the evolutionary history of SARS-CoV-2 is largely unknown. A recent analysis of genetic information and the spike (S) protein structure(4, 5) highlights two notable features of the SARS-CoV-2 genome. First, the receptor binding domain (RBD) of SARS-CoV-2 is distinct from the most closely-related virus (RaTG13) of bat origin and more closely related to pangolin coronaviruses(6, 7). The spike protein of SARS-CoV-2 is demonstrated to have a high affinity for the human ACE2 receptor molecule(4). Second, a unique insertion of 12 nucleotides (encoding four amino acids, PRRA) at the S1/S2 boundary(8) leading to a predictively solvent-exposed PRRAR|SV sequence, which corresponds to a canonical furin-like cleavage site(9, 10).

With respect to the first feature, an RBD identified in a SARS-like virus from a pangolin suggests that an RBD similar to that of SARS-CoV-2 may already exist in mammalian host(s) prior to its introduction into humans(7). The question remaining is the history and function of the insertion at the S1/S2 boundary, which is unique to SARS-CoV-2. By sequencing the whole genome of SARS-CoV-2 from cell isolates and clinical samples, we identified two deletion variants that directly affect the furin cleavage site itself (NSPRRAR) or a flanking sequence (QTQTN). We screen these two deletions in cell-isolated strains and clinical samples. To explore the potential effect of these deletions, these two deletion variants were isolated and their replication kinetics were investigated in both Vero and Vero-E6 cells.

Results

Identification of deletions in SARS-CoV-2 spike protein

The first COVID-19 clinical case (Sample 014, Table1) in Guangdong was reported on 19th January, with illness onset on 1st January(11). A BALF (bronchoalveolar lavage fluid) sample from this patient was collected and inoculated on Vero-E6 cells. A cell-isolated viral strain was obtained after three rounds of passage. Multiple sequencing methods were used for whole genome sequencing and the validation of variants (Figure1 A, Table1), including multiplex-PCR with Miseq platform (PE150), direct CDNA sequencing in Nanopore platform and Sanger sequencing (See Materials and Methods for detail). After mapping to the SARS-CoV-2 reference genome (MN908947.3), we found that there were two variants in the cell-isolated viral strain with deletions at (1) 23585–23599 (Var1), flanking the polybasic cleavage site, resulting in a QTQTN deletion in the spike protein (one amino acid before the polybasic cleavage site) and (2) 23597–23617 (Var2), resulting in a NSPRRAR deletion that includes the polybasic cleavage site (Figure 1A). To exclude the possibility that these findings were caused by errors in PCR amplification, both of the deletion variants were verified through direct cDNA sequencing on the ONT nanopore platform. Sanger sequencing with specific primers also identified heterozygous peaks with distinct double peaks starting at the position 23585 and triple peaks after that, highlighting the existence of multiple variants caused by the above two deletions (Figure 1B). To investigate the dynamics of these deletion variants, we performed nanopore sequencing on the 014 viral strain, isolated at different rounds of passage from the Vero-E6 cell culture (Figure 1C). High frequencies of the deletion variant Var1were observed after the first passage and high frequencies of the deletion variant Var2 were observed after the 4th passage, at which point the frequency of Var1 and Var2 reached around 50%. The percentages of these two deletion variants were steady in the following

passages.

The deletion is commonly identified in cell isolated strains

To investigate whether the deletions described above were random mutations that occasionally arise in a strain, or whether they commonly occur after cell passages, we performed whole genome sequencing on 23 other SARS-CoV-2 strains collected after two rounds of cell passage in Vero-E6 or Vero cells (Table 1). The corresponding original samples for these strains were collected between 19th January and 28th February 2020. In addition to the 014 strain mentioned above, 10 out of 18 Vero-E6 isolated strains and 1 out of 5 Vero isolated strains displayed the Var1 deletion variant (>10% of sequencing reads; Figure 1D). Additionally, in two Vero-E6 isolated strains (619 and 4276), Var2 was detected, and this variant has been independently identified by another group almost at the same time, using direct RNA sequencing method(12). To find out whether these deletions were restricted to a specific genetic lineage, we next investigated the phylogenetic relationship of these viral strains. As shown in Figure 1D, the strains with a relatively higher ratio of this deletion were dispersed in the phylogenetic tree, that suggesting the deletion mutations did not arise through shared ancestry and were not restricted to a specific genetic lineage of SARS-CoV-2 viruses.

Replication kinetics of the deletion variants

To evaluate the effect of these deletions on virus replication, we performed plaque assays and picked individual clones for different variants. Single plaques for Var1 and Var2 were obtained and confirmed by whole genome sequencing (014-Var1, 014-Var2; Table 1). However, the 014 strain without these deletions could not be successfully selected from plaques, possibly due to the replication advantage of

the deletion variants in cell culture. We investigated the replication kinetics of 014-Var1 and 014-Var2 in Vero-E6 and Vero cells. The strain 029/E6 was used as a reference, which has no deletion mutations and only one amino acid difference from strain 014 on the spike protein (H47Y). The viral replication kinetics were assessed by detecting the intracellular viral loads at 1, 3, 6, 9, 12 and 24 hours post inoculation (Figure 2). As shown in Figure 2A, the 014-Var1 and 014-Var2 exhibit similar replication dynamics to the 029 strain in Vero-E6 cells. In contrast, the deletion of 23583–23599 in SARS-CoV-2 (Var1) significantly diminishes cellular viral load at 24 hours post-inoculation in Vero cells (Figure 2B) and to a lesser extent in Vero-E6 cells (Figure 2A). This is the possible reason that 014-Var1 was observed less often in Vero cells than in Vero-E6 cells (Figure 1D).

Screening for deletion variants in original clinical samples

To identify whether these deletions also occurred in the original clinical samples, we screened high-throughput sequencing data from 149 clinical samples, which were collected between 6th February and 20th March in Guangdong, China. There were 68 SARS-CoV-2 genomes, with an average sequencing depth ≥20 at the sites neighboring 23585. As shown in Table 2, variants with the QTQTN (Var1) were found in 3 (4%) of clinical samples, with the ratio of deletion variant in total reads ranging from 8.8–32.8%, indicating that this deletion also occurs in *in vivo* infections. Notably, two out of the three patients from which these samples were derived displayed mild symptoms and recurrence of SARS-CoV-2 infection after being discharged from hospital. The sequenced samples were collected at 4 days and 17 days after discharge, respectively. The third case (20SF5645) was an asymptomatic infection case. To date, there are no genome sequences deposited in public databases containing these two deletions. While the described Var1 deletion variant was only detected in clinical samples after

deep sequencing, such variants may be underrepresented in databases due to the low frequency and consequent elimination upon consensus sequence generation.

Discussion

The spike protein of coronaviruses plays an important role in viral infectivity, transmissibility and antigenicity. Therefore, the genetic character of the spike protein in SARS-CoV-2 may shed light on its origin and evolution(7, 8). For SARS-CoV-1, positive selection was identified in the spike coding sequence(13) and deletions in *ORF8*(14) during the early, but not late, stage of the epidemic, suggesting that SARS-CoV-1 may have been sub-optimal in the human population during the early epidemic stage after it was first transmitted from an intermediate animal host, and underwent further adaptation. SARS-CoV-2, however, has presented high infectivity and efficient transmission capability since its identification(1) suggesting the polybasic cleavage site is an important component of the virus' fitness within the human population. Genetic changes related to viral fitness of SARS-CoV-2 require further epidemiological investigation and functional analysis.

Here, we use different sequencing methods to identify and verify two deletion variants either directly affecting the polybasic cleavage site (Var1) or a site immediately upstream of it (Var2). The QTQTN deletion variant (Var1) was detected in 3 out of 68 clinical samples and half of the 24 *in vitro* isolated viral strains tested in this study. The cellular replication kinetic data suggests the deletion of the polybasic cleavage site does not affect SARS-CoV-2 replication in Vero and Vero-E6 cells, whilst the QTQTN deletion may restrict virus replication in Vero cells at the late phase. These data indicate that (i) the deletions of QTQTN and the polybasic cleavage site are likely under strong purifying selection *in*

vivo, since the deletion is rarely identified in clinical samples, (ii) there may be an efficient mechanism for generating these deletions, given that the QTQTN deletion (Var1) is commonly detected after two rounds of cell passage and (iii) the PRRA insertion, which distinguishes SARS-CoV-2 from other SARS-like viruses, is not fixed *in vitro*, because the NSPRRAR deletion variant (Var2) is observed in 3 out of 24 Vero-E6 isolated strains, but does appear to be subject to purifying selection *in vivo*.

Given that these residues are located in solvent-accessible loops of the spike protein, combined with the observation that they are either partially (QTQTN) or completely (NSPRRAR) unresolved in recently reported SARS-CoV-2 S cryoEM structures(4, 5) (Figure 3), it seems likely that this region is structurally tolerant to deletions. Whilst the deletion of the furin site, as observed in Var2, would result in a loss of susceptibility to furin cleavage at this site, the effect of Var1 on furin cleavage is less evident. However, it is likely that these overlapping deletion variants have arisen through the same selective pressure and are therefore both likely to compromise furin-mediated cleavage at this position in the S protein, albeit possibly to different extents. Furthermore, it is possible that the presence of a conserved cathepsin L site 10 residues downstream of the polybasic cleavage site may provide functional tolerance(15) to any reduction in proteolytic cleavage efficiency that may arise from changes in this region (Figure 1A). Consistent with the modeling analysis, the replication dynamics in Vero and Vero-E6 cells also indicate that polybasic cleavage site deletion (Var2) does not affect virus replication *in vitro*.

Notably, a recently reported SARS-like strain, RmYN02, which is phylogenetically related to SARS-CoV-2, also has a possible deletion at the QTQT site(16). This raises another possible scenario,

which is that some SARS-CoV-2-like viruses in animals may not have had QTQTN in their spike protein. The origin of polybasic cleavage site (PRRA) is important to understanding the evolution history and tracing the potential animal reservoir(s) of SARS-CoV-2. Here, the different deletion frequencies observed *in vitro* and *in vivo* have provide clues that will aid further investigation of this evolutionary tale. The absence of NSPRRA in isolated SARS-CoV-2 strains could be used to further investigate its infectivity in different potential intermediate animal hosts and resolve the origin of this feature of the SARS-CoV-2 genome. In addition, the different selective pressure observed on NSPRRA region of SARS-CoV-2 *in vivo* and *in vitro* highlight the NSPRRA deletion variant generated in this study as a promising vaccine candidate in the future.

Materials and Methods

Ethics

This study was approved by ethics committee of the Center for Disease Control and Prevention of Guangdong Province. Written consent was obtained from patients or their guardian(s) when clinical samples were collected. Patients were informed about the surveillance before providing written consent, and sequence data were analyzed anonymously.

Viral isolation

Vero E6 or Vero cells were used for SARS-CoV-2 virus isolation and passage. The cells were inoculated with 100 µl processed patient sample. Cytopathic effect (CPE) were observed daily. If there was no CPE observed, cell lysis was collected by centrifugation after three repeated freeze-thaw and 100 µl supernatant were used for the second round of passage.

Genetic sequencing and sequence analysis

The deletion variants of SARS-CoV-2 were confirmed by different approaches as previously described(17) (i) using version 1 of the ARTIC COVID-19 multiplex PCR primers (https://artic.network/ncov-2019), followed by sequencing on a Miseq PE150 or an ONT MinION, (ii) CDNA directly sequencing on an ONT MinION and (iii) sanger sequencing by using the nCoV-2019_78_LEFT and nCoV-2019_78_RIGTH primers from the ARTIC COVID-19 multiplex PCR primers set. The amplification products targeting the 23444-23823 fragment of viral genome (numbered according to MN908947.3).

For metatranscriptomics, total RNAs were extracted from different types of samples by using QIAamp Viral RNA Mini Kit, followed by DNase treatment and purification with TURBO

DNase and Agencourt RNAClean XP beads. Libraries were prepared using the SMARTer Stranded Total RNA-Seq Kit v2 (according to the manufacturer's protocol starting with 10 ng total RNA. Sequencing of metatranscriptome libraries was conducted on the Illumina Miseq PE 150 platform. For the multiplex PCR approach, we followed the general method of multiplex PCR as described in (https://artic.network/ncov-2019)(18). Briefly, multiplex PCR was performed with two pooled primer mixtures and cDNA reverse-transcribed with random primers was used as a template. After 25-35 rounds of amplification, PCR products were collected and quantified, followed by sequencing on Illumina Miseq PE 150 platform or MinION sequencing device. Assembly of the Illumina raw data was performed using Geneious v11.0.3 (https://www.geneious.com). Assembly of the nanopore raw data was performed using the ARTIC bioinformatic pipeline for COVID-19 with minimap2(19) and medaka (https://github.com/nanoporetech/medaka) for consensus sequence generation. Variant sites were called by using iVar(20) with depth >=20 as a threshold. For direct cDNA sequencing, we followed the Nanopore Direct cDNA sequencing protocol (SQK-DCS109). Briefly, 100ng viral RNA were reverse transcripted using SuperScript™ IV First-Strand Synthesis System (Invitrogen, USA) followed by RNA chain digestion and second strand synthesis. A total of 20ng cDNA libraries were loaded to FLO-MIN106 flow cell. Generated sequences were mapped to MN908947.3 reference sequence using minimap2. The ML phylogeny for 24 viral strains genomes was estimated with PhyML(21) using the HKY+Γ_4 substitution model(22) with gamma-distributed rate variation(23).

Viral kinetics analysis

The individual clones of deletion variants were selected by using a plaque assay. The isolated 014 strains were serially-diluted and used to inoculate the monolayer of Vero-E6 cells. When CPE were observed, the cell monolayers were scraped with the back of a pipette tip. Virus lysate was used for genetic sequencing and viral strain amplification. To assess the kinetic of virus replication, different viral strains were first tiltered and inoculated with Vero-E6 and Vero cells at MOI 0.5. Time was set as zero when cells were incubated with viruses. After 1 hour adsorption, the culture media were removed and cells were washed twice with PBS to remove unattached virus. Cells were lysed at different time post inoculation and total RNA was extracted by using RNeasy mini kit (QIAGEN, Germany). Cellular viral loads were calculated by using SARS-CoV-2 RT-PCR kit (DAAN GENE, Guangzhou, China) and GAPDH (glyceraldehyde-3-phosphate dehydrogenase) gene was parallelly quantified as an endogenous control.

Data Availability

Metagenomic sequencing, multiplex PCR sequencing and cDNA direct sequencing data after mapping to SARS-COV-2 reference genome (MN908947.3) have been deposited in the Genome Sequence Archive(24) in BIG Data Center(25), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under project accession numbers CRA002500, publicly accessible at https://bigd.big.ac.cn/gsa. The sample information and corresponding accession number for each sample are listed in the Table 1.

Acknowledgments

This work was supported by grants from Guangdong Provincial Novel Coronavirus Scientific and

259 Technological Project (2020111107001), Science and Technology Planning Project of Guangdong
260 (2018B020207006). The Welcome Centre for Human Genetics is supported by Wellcome Centre grant
261 203141/Z/16/Z.

262 Conflict of interest: None declared.

263 **Reference**

264 **1. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y,**
265 **Yuan M-L, Zhang Y-L, Dai F-H, Liu Y, Wang Q-M, Zheng J-J, Xu L, Holmes EC, Zhang**
266 **Y-Z. 2020. A new coronavirus associated with human respiratory disease in China. Nature**
267 **1–8.**

268 **2. WHO (2020). Coronavirus disease (COVID-2019) situation reports.**

269 **3. Cui J, Li F, Shi Z-L. 2019. Origin and evolution of pathogenic coronaviruses. Nat Rev**
270 **Microbiol 17:181–192.**

271 **4. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. 2020. Structure,**
272 **Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell.**

273 **5. Li F. 2005. Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed**
274 **with Receptor. Science 309:1864–1868.**

275 **6. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, Li N, Guo Y, Li X, Shen X, Zhang Z,**
276 **Shu F, Huang W, Li Y, Zhang Z, Chen R-A, Wu Y-J, Peng S-M, Huang M, Xie W-J, Cai**
277 **Q-H, Hou F-H, Chen W, Xiao L, Shen Y. 2020. Isolation of SARS-CoV-2-related**
278 **coronavirus from Malayan pangolins. Nature 1–4.**

279 **7. Lam TT-Y, Shum MH-H, Zhu H-C, Tong Y-G, Ni X-B, Liao Y-S, Wei W, Cheung WY-M, Li**
280 **W-J, Li L-F, Leung GM, Holmes EC, Hu Y-L, Guan Y. 2020. Identifying SARS-CoV-2**
281 **related coronaviruses in Malayan pangolins. Nature.**

8. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. Nat Med.

9. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Research 176:104742.

10. Izaguirre G. 2019. The Proteolytic Regulation of Virus Cell Entry by Furin and Other Proprotein Convertases. Viruses 11:837.

11. Kang M, Wu J, Ma W, He J, Lu J, Liu T, Li B, Mei S, Ruan F, Lin L, Zou L, Ke C, Zhong H, Zhang Y, Chen X, Liu Z, Zhu Q, Xiao J, Yu J, Hu J, Zeng W, Li X, Liao Y, Tang X, Xiao S, Wang Y, Song Y, Zhuang X, Liang L, Zeng S, He G, Lin P, Deng H, Song T. 2020. Evidence and characteristics of human-to-human transmission of SARS-CoV-2. medRxiv 2020.02.03.20019141.

12. Davidson AD, Williamson MK, Lewis S, Shoemark D, Carroll MW, Heesom K, Zambon M, Ellis J, Lewis PA, Hiscox JA, Matthews DA. 2020. Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site. bioRxiv 2020.03.22.002204.

13. The Chinese SARS Molecular Epidemiology Consortium. 2004. Molecular Evolution of the SARS Coronavirus During the Course of the SARS Epidemic in China. Science 303:1666–1669.

14. Muth D, Corman VM, Roth H, Binger T, Dijkman R, Gottula LT, Gloza-Rausch F, Balboni A, Battilani M, Rihtarič D, Toplak I, Ameneiros RS, Pfeifer A, Thiel V, Drexler JF, Müller MA, Drosten C. 2018. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. Scientific Reports 8:1–11.

15. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu N-H, Nitsche A, Müller MA, Drosten C, Pöhlmann S. 2020. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell S0092-8674(20)30229–4.

16. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, Wang P, Liu D, Yang J, Holmes EC, Hughes AC, Bi Y, Shi W. 2020. A novel bat coronavirus reveals natural insertions at the S1/S2 cleavage site of the Spike protein and a possible recombinant origin of HCoV-19. bioRxiv 2020.03.02.974139.

17. Lu J, du Plessis L, Liu Z, Hill V, Kang M, Lin H, Sun J, François S, Kraemer MUG, Faria NR, McCrone JT, Peng J, Xiong Q, Yuan R, Zeng L, Zhou P, Liang C, Yi L, Liu J, Xiao J, Hu J, Liu T, Ma W, Li W, Su J, Zheng H, Peng B, Fang S, Su W, Li K, Sun R, Bai R, Tang X, Liang M, Quick J, Song T, Rambaut A, Loman N, Raghwani J, Pybus OG, Ke C. 2020. Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell S0092867420304864.

18. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, Oliveira G, Robles-Sikisaka R, Rogers TF, Beutler NA, Burton DR, Lewis-Ximenez LL, de Jesus JG,

Giovanetti M, Hill SC, Black A, Bedford T, Carroll MW, Nunes M, Jr LCA, Sabino EC, Baylis SA, Faria NR, Loose M, Simpson JT, Pybus OG, Andersen KG, Loman NJ. 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nature Protocols 12:1261–1276.

19. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.

20. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, Tan AL, Paul LM, Brackney DE, Grewal S, Gurfield N, Van Rompay KKA, Isern S, Michael SF, Coffey LL, Loman NJ, Andersen KG. 2019. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biology 20:8.

21. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321.

22. Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174.

23. Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314.

24. Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, Tang B, Dong L, Ding N, Zhang Q, Bai Z, Dong X, Chen H, Sun M, Zhai S, Sun Y, Yu L, Lan L, Xiao J, Fang X, Lei H, Zhang Z,

Zhao W. 2017. GSA: Genome Sequence Archive. Genomics Proteomics Bioinformatics 15:14–18.

25. National Genomics Data Center Members and Partners. 2020. Database Resources of the National Genomics Data Center in 2020. Nucleic Acids Res 48:D24–D33.

Figure legends

Figure1. Deletion variants identified in SARS-CoV-2 cell strains. (A) High-throughput sequencing of the cell-isolated strain (014) from the first SARS-CoV-2 patient (EPI 403934) in Guangdong, China. Representative reads mapping to the SARS-CoV-2 genome (MN908947.3 used as reference genome) showed two deletion variants. Redundant proteolytic cleavage sites including furin cleavage site (PRRARS|V) and cathepsin L site (QSIIAY|T) are marked with red arrows (B) Sanger sequencing of the 014 cell strains. Heterozygous peaks are highlighted with a red box and sites with distinct three peaks are marked with * (C) Results of high-throughput sequencing, showing the ratio of deletion variants in original clinical sample SF014 (P0) and in cell strains, after 7 rounds of cell passage (P1-7). The size of square was proportion to the number of reads having these deletions. (D) Phylogenetic tree of genome sequences of all 24 SARS-CoV-2 cell strains (see Table 1). The size of the circles is proportional to the percentage of Var1 (QTQTN deletion at 23585–23599) in total reads, except for strains 619, 4279 and 014 in which Var2 deletions were detected. The maximum likelihood tree was rooted with the reference genome MN908947.3.

Figure 2. The replication kinetics of the deletion variants in Vero-E6 and Vero cells. Vero-E6 and Vero cells were infected with the isolated strains 014_Var1, 014_Var2, and 029/E6 (Table 1) at multiplicity of infection (MOI) 0.5. Viral RNA was quantified by real-time PCR using GAPDH as endogenous control. At the each time point, the relative fold-change in total intracellular viral RNA was measured by comparison with the viral RNA level at 1-hour post inoculation. Data are the mean ± SD of three independent experiments. Asterisk indicate the significant difference ($p<0.05$).

Figure3. Observed deletions near the S1/S2 boundary map to a unresolved region in the cryoEM structure of SARS-CoV-2 S. Cartoon representation of the SARS-CoV-2 S protein ectodomain, as resolved by Walls and colleagues(4) (PDB: 6VXX). The S1 and S2 subunits of the different protomers are indicated (white and grey, respectively). The unresolved loop that contains part of deletion Var1 (^{675}QTQTN^{679}) and all of deletion Var2 (^{679}NSPRRAR^{685}) is indicated within each protomer of the trimeric assembly through signposting flanking residues T^{676} and S^{689} as spheres in deep teal. Similarly, the first residue of Var1 (Q^{675}), which is resolved in the structure, is indicated as an orange surface within each of the S protomers. N-linked glycans are shown as blue spheres and the Asn side chains to which the glycans are linked are presented as sticks. Inset: A zoomed-in side view representation of this local arrangement is shown. T^{676} and S^{689}, which flank the unresolved loop, and Var1 residue Q^{675} are numbered and indicated under transparent spheres as deep teal and orange sticks, respectively. A dashed line indicating the approximate position of the connecting unresolved loop is shown. N-linked glycans are presented as in the original image with their residue numbers marked.

385 **Table1. Sample information and accession numbers for all sequences**

Patient identifier	Sample isolated from	Passage	Sample name	Sequencing method	Accession number
Case1	BALF	Original	014	Metagenomic	SAMC151281
	Vero-E6	3	014/MiSeq	PCR+MiSeq	SAMC150996
	Vero-E6	3	014/cDNA	Nanopore direct cDNA	SAMC150997
	Vero-E6	Plaque	014_Var1	PCR+Nanopore	SAMC192628
	Vero-E6	Plaque	014_Var2	PCR+Nanopore	SAMC192629
Case2	Vero-E6	2	025/E6	PCR+Nanopore	SAMC150991
Case3	Vero	2	028/Vero	PCR+Nanopore	SAMC150988
	Vero-E6	2	028/E6	PCR+Nanopore	SAMC150992
Case4	Vero-E6	2	029/E6	PCR+Nanopore	SAMC150975
Case5	Vero-E6	2	107/E6	PCR+Nanopore	SAMC150977
	Vero	2	107/Vero	PCR+Nanopore	SAMC150989
Case6	Vero-E6	2	108/E6	PCR+Nanopore	SAMC150993
	Vero	2	108/Vero	PCR+Nanopore	SAMC150995
Case7	Vero-E6	2	112/E6	PCR+Nanopore	SAMC150976
	Vero	2	112/Vero	PCR+Nanopore	SAMC150994
Case8	Vero-E6	2	115/E6	PCR+Nanopore	SAMC150978
	Vero	2	115/Vero	PCR+Nanopore	SAMC150990

Case9	Vero-E6	2	252/E6	PCR+Nanopore	SAMC150980
Case10	Vero-E6	2	262/E6	PCR+Nanopore	SAMC150981
Case11	Vero-E6	2	263/E6	PCR+Nanopore	SAMC150983
Case12	Vero-E6	2	265/E6	PCR+Nanopore	SAMC150982
Case13	Vero-E6	2	272/E6	PCR+Nanopore	SAMC150984
Case14	Vero-E6	3	619/E6	PCR+Nanopore	SAMC153235
Case15	Vero-E6	2	1676/E6	PCR+Nanopore	SAMC150979
Case16	Vero-E6	3	4276/E6	PCR+Nanopore	SAMC153234
Case17	Vero-E6	2	F2/E6	PCR+Nanopore	SAMC150985
Case18	Vero-E6	2	F4/E6	PCR+Nanopore	SAMC150986
Case19	Vero-E6	2	F5/E6	PCR+Nanopore	SAMC150987
Case20	nasopharyngeal	Original	20SF5645	PCR+Nanopore	SAMC150972
Case21	nasopharyngeal	Original	ST-N3-D	PCR+Nanopore	SAMC150973
Case22	nasopharyngeal	Original	SZ-N16-D	PCR+Nanopore	SAMC150974

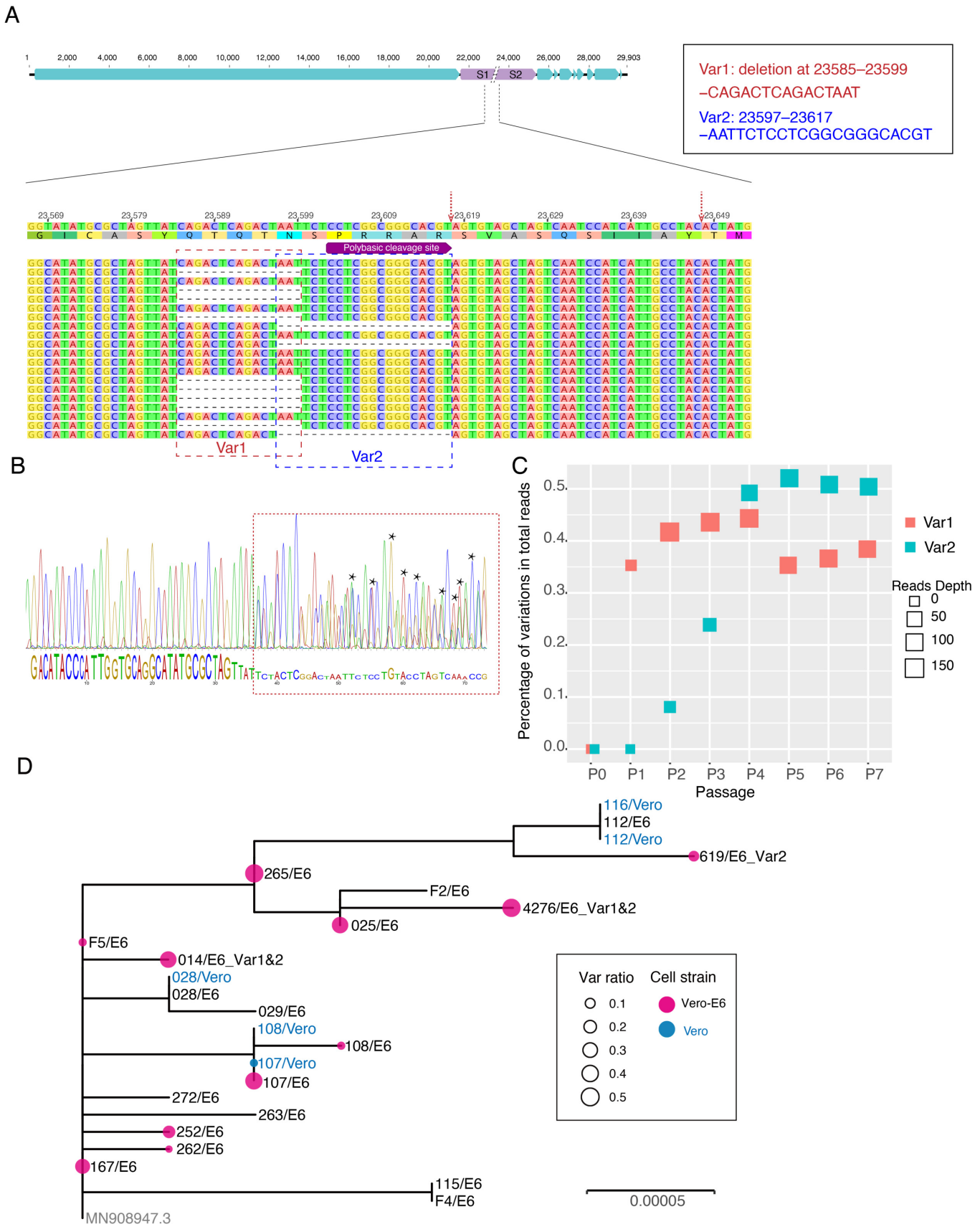
386

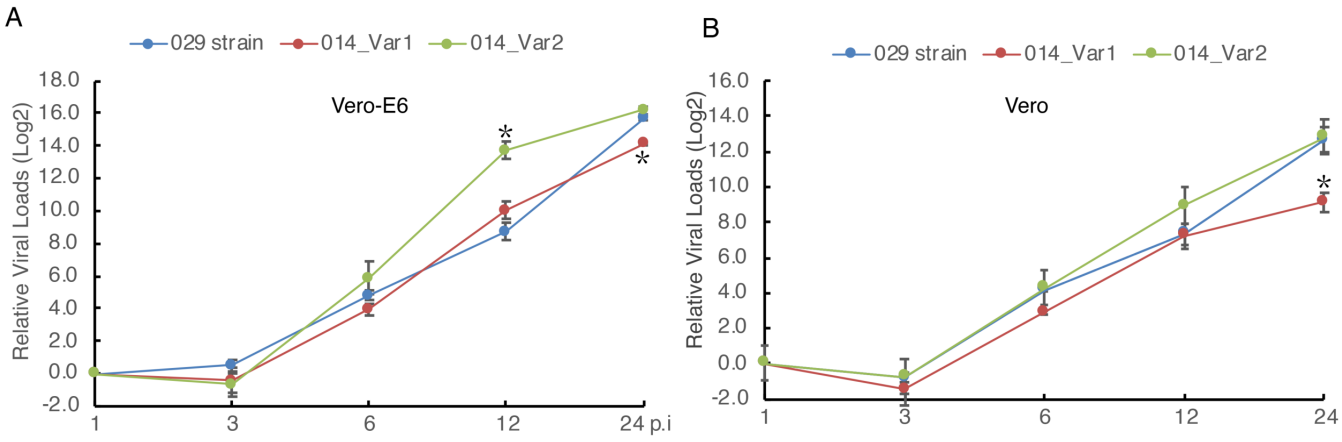
387

388 **Table 2: QTQTN deletion variant (23585–23599, Var1) identified in clinical samples**

Samples	Days post illness onset	REF_depth	ALT_depth	Del Variant Ratio
20SF5645	Asymptomatic	104	25	19.4%
ST-N3-D*	16	82	8	8.8%
SZ-N16-D*	30	256	125	32.8%

389 * Cases detected with the recurrence of SARS-CoV-2 after discharge





Journal of Virology

JVI

Journal of Virology