Identification of Common Deletions in the Spike Protein of Severe Acute Respiratory Syndrome Coronavirus 2

Zhe Liu,[a,b] Huanying Zheng,[b] Huifang Lin,[a,b] Mingyue Li,[c] Runyu Yuan,[a,b] Jinju Peng,[a,b] Qianling Xiong,[a,b] Jiufeng Sun,[a,b] Baisheng Li,[b] Jie Wu,[b] Lina Yi,[a,b] Xiaofang Peng,[a,b] Huan Zhang,[a,b] Wei Zhang,[a,b] Ruben J. G. Hulswit,[d] Nick Loman,[e] Andrew Rambaut,[f] Changwen Ke,[b] Thomas A. Bowden,[d] Oliver G. Pybus,[g,h] Jing Lu[a,b]

[a]Guangdong Provincial Institution of Public Health, Guangzhou, China
[b]Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China
[c]Department of Rehabilitation Medicine, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, China
[d]Division of Structural Biology, Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom
[e]Institute of Microbiology and Infection, University of Birmingham, Birmingham, United Kingdom
[f]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom
[g]Department of Zoology, University of Oxford, Oxford, United Kingdom
[h]Department of Pathobiology and Population Sciences, The Royal Veterinary College, London, United Kingdom

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

ABSTRACT Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus first identified in December 2019. Notable features that make SARS-CoV-2 distinct from most other previously identified betacoronaviruses include a receptor binding domain and a unique insertion of 12 nucleotides or 4 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 12 of 24 *in vitro*-isolated viruses, while 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.

IMPORTANCE The spike protein determines the infectivity and host range of coronaviruses. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has two unique features in its spike protein, the receptor binding domain and an insertion of 12 nucleotides at the S1/S2 boundary resulting in 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 we 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 sequence and the flanking QTQTN sequence 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.

Citation Liu Z, Zheng H, Lin H, Li M, Yuan R, Peng J, Xiong Q, Sun J, Li B, Wu J, Yi L, Peng X, Zhang H, Zhang W, Hulswit RJG, Loman N, Rambaut A, Ke C, Bowden TA, Pybus OG, Lu J. 2020. Identification of common deletions in the spike protein of severe acute respiratory syndrome coronavirus 2. J Virol 94:e00790-20. https://doi.org/10.1128/JVI.00790-20.

Editor Julie K. Pfeiffer, University of Texas Southwestern Medical Center

Copyright © 2020 American Society for Microbiology. All Rights Reserved.

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

Received 26 April 2020
Accepted 16 June 2020
Accepted manuscript posted online 22 June 2020
Published 17 August 2020

KEYWORDS COVID-19, SARS-CoV-2, deletion mutation, replication kinetics, spike protein

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus that was first identified at the end of December 2019 (1) and is responsible for the global pandemic of COVID-19 (2). Unlike two other zoonotic coronaviruses, SARS-CoV-1 and Middle East respiratory syndrome (MERS)-CoV (3), the evolutionary history of SARS-CoV-2 is largely unknown. Recent analyses of genetic information and the spike (S) protein structure (4, 5) highlight 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 is 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) leads to a predicted 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 could detect these two deletions in both 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, Table 1) in Guangdong, China, was reported on 19 January, with illness onset on 1 January (11). A bronchoalveolar lavage fluid (BALF) 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 (Fig. 1A; Table 1), including multiplex PCR with the MiSeq platform (PE150), direct cDNA sequencing with the Nanopore platform, and Sanger sequencing (see Materials and Methods for details). After mapping sequences to the SARS-CoV-2 reference genome (GenBank accession number MN908947.3), we found that there were two variants in the cell-isolated viral strain with deletions at (i) positions 23583 to 23597 (Var1), flanking the polybasic cleavage site, resulting in a QTQTN deletion in the spike protein (one amino acid before the polybasic cleavage site); and (ii) positions 23597 to 23617 (Var2), resulting in a NSPRRAR deletion that includes the polybasic cleavage site (Fig. 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 Oxford Nanopore Technologies (ONT) platform. Sanger sequencing with specific primers also identified heterozygous peaks, with distinct double peaks starting at the position 23583 and triple peaks after that, highlighting the existence of multiple variants caused by the above two deletions (Fig. 1B). To investigate the dynamics of these deletion variants, we performed nanopore sequencing on the 014 viral strain, isolated at different rounds of passage on Vero-E6 cells (Fig. 1C). High frequencies of the deletion variant Var1 were 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.

TABLE 1 Sample information and accession numbers for all sequences

Patient identifier	Sample isolated from:	Passage	Sample name	Sequencing method	Genome Sequence Archive accession no.
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 swab	Original	20SF5645	PCR+Nanopore	SAMC150972
Case21	Nasopharyngeal swab	Original	ST-N3-D	PCR+Nanopore	SAMC150973
Case22	Nasopharyngeal swab	Original	SZ-N16-D	PCR+Nanopore	SAMC150974

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 19 January and 28 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) (Fig. 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 a 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 Fig. 1D, the strains with a relatively higher ratio of this deletion were dispersed in the phylogenetic tree, suggesting that 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, 12, and 24 hours

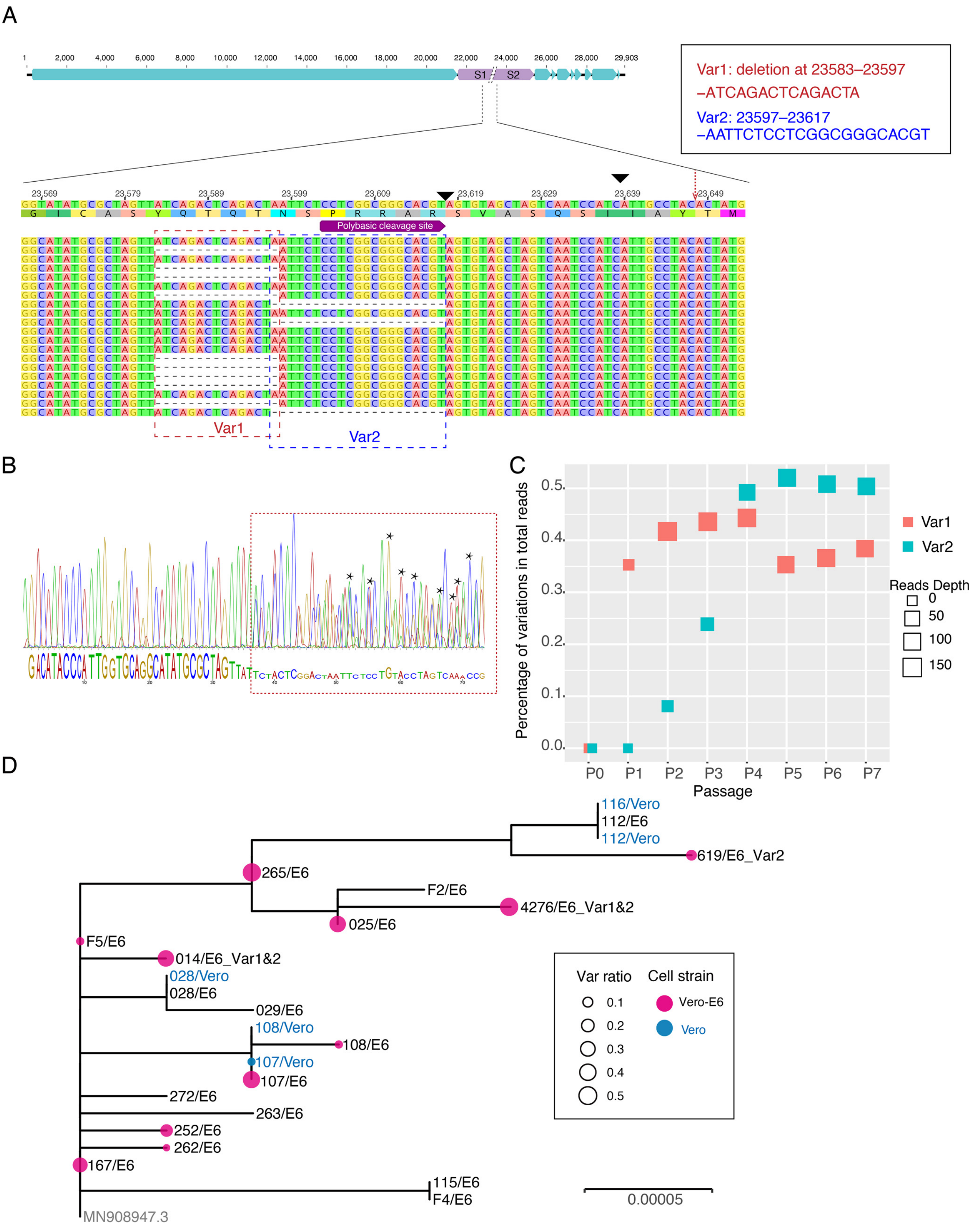

FIG 1 Deletion variants identified in SARS-CoV-2 cell strains. (A) High-throughput sequencing of the cell-isolated strain (014) from the first SARS-CoV-2 patient (EPI 403934) in Guangdong, China. Representative reads mapping to the SARS-CoV-2 genome (GenBank accession number MN908947.3 used as reference genome) showed two deletion variants. Redundant proteolytic cleavage sites, including furin cleavage site (PRRARS|V) and cathepsin L site (QSIIAY|T) are

(Continued on next page)

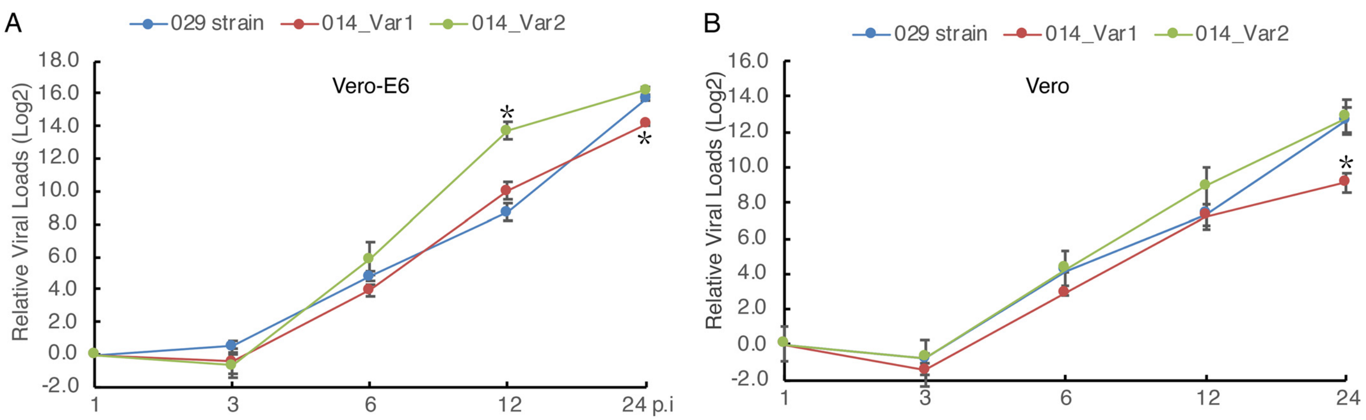

FIG 2 The replication kinetics of the deletion variants in Vero-E6 and Vero cells. Vero-E6 and Vero cells were infected with the isolated strains 014_Var1, 014_Var2, and 029/E6 (Table 1) at a multiplicity of infection (MOI) of 0.5. Viral RNA was quantified by real-time PCR using GAPDH as an endogenous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

postinoculation (Fig. 2). As shown in Fig. 2A, 014-Var1 and 014-Var2 exhibit similar replication dynamics to the 029 strain in Vero-E6 cells. In contrast, the deletion of positions 23583 to 23597 in SARS-CoV-2 (Var1) significantly diminishes cellular viral load at 24 hours postinoculation in Vero cells (Fig. 2B) and to a lesser extent in Vero-E6 cells (Fig. 2A). This is a possible reason that 014-Var1 was observed less often in Vero cells than in Vero-E6 cells (Fig. 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 6 February and 20 March in Guangdong, China. There were 68 SARS-CoV-2 genomes, with an average of ≥20× sequencing depth at the sites neighboring 23583. As shown in Table 2, variants with QTQTN (Var1) were found in 3 (4%) of the clinical samples, with the ratio of deletion variants in total reads ranging from 8.8% to 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 the 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 suboptimal in the human population during the early epidemic stage after

FIG 1 Legend (Continued)

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 to P7). The size of each square was proportional 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 positions 23583 to 23597) 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 sequence under GenBank accession number MN908947.3.

TABLE 2 QTQTN deletion variant[a] identified in clinical samples

Sample	Days post-illness onset	REF_depth (×)[c]	ALT_depth (×)[c]	Del variant ratio (%)
20SF5645	Asymptomatic	104	25	19.4
ST-N3-D[b]	16	82	8	8.8
SZ-N16-D[b]	30	256	125	32.8

[a] At position 23583 to 23597 (Var1).
[b] Cases detected with the recurrence of SARS-CoV-2 after discharge.
[c] REF_depth, sequencing depth for the reference bases; ALT_depth, sequencing depth for the altered bases.

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 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 used different sequencing methods to identify and verify two deletion variants directly affecting either 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 in 12 of the 24 *in vitro*-isolated viral strains tested in this study. The cellular replication kinetics data suggest that deletion of the polybasic cleavage site does not affect SARS-CoV-2 replication in Vero and Vero-E6 cells, while the QTQTN deletion may restrict virus replication in Vero cells at the late phase. These data indicate that (i) the QTQTN and the polybasic cleavage site sequences 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 it does appear to be subject to purifying selection *in vivo*.

Given that these residues are located in solvent-accessible loops of the spike protein, and that they are either partially (QTQTN) or completely (NSPRRAR) unresolved in recently reported SARS-CoV-2 S cryo-electron microscopy (EM) structures (4, 5) (Fig. 3), it seems likely that this region is structurally tolerant to deletions. While 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 (Fig. 1A). Consistent with this hypothesis, the replication dynamics in Vero and Vero-E6 cells also indicate that polybasic cleavage site deletion (Var2) does not affect virus replication *in vitro*.

Remarkably, the recently reported SARS-CoV-2-like bat strain, RmYN02, also displays a deletion of QTQT (16), indicating that some SARS-CoV-2-like viruses in animals may not encode QTQTN in their spike gene. The origin of the polybasic cleavage site (PRRA) is important for understanding the evolutionary history and for tracing the potential animal reservoir(s) of SARS-CoV-2. Here, the different deletion frequencies observed *in vitro* and *in vivo* have provided 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 pressures observed on the NSPRRA region of SARS-CoV-2 *in vivo* and *in vitro*

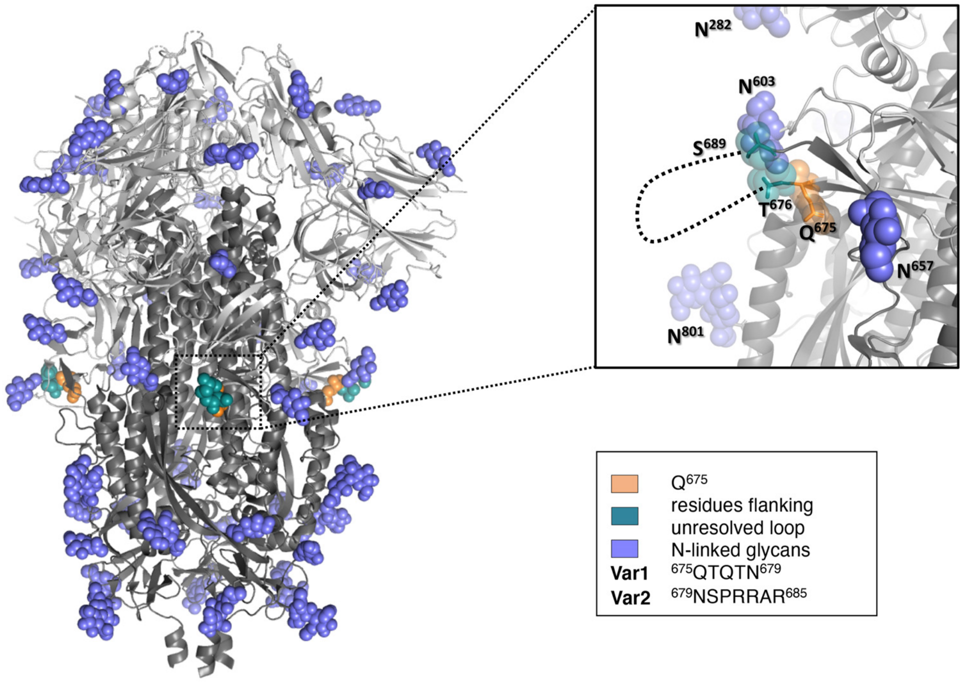

FIG 3 Observed deletions near the S1/S2 boundary map to an unresolved region in the cryo-EM structure of SARS-CoV-2 S. Cartoon representation of the SARS-CoV-2 S protein ectodomain, as resolved by Walls and colleagues (4) (PDB ID 6VXX). The S1 and S2 subunits of the different protomers are indicated (white and gray, 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 colored orange within each of the S protomers. N-linked glycans are shown as blue spheres and the Asn side chains to which the glycans that 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.

highlight the NSPRRA deletion variant observed in this study as a promising vaccine candidate in the future.

MATERIALS AND METHODS

Ethics. This study was approved by ethics committee of the Guangdong Provincial Center for Disease Control and Prevention. 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 isolation and passage. The cells were inoculated with 100-μl processed patient sample. Cytopathic effect (CPE) was observed daily. If there was no CPE observed, cell lysates were collected by centrifugation after three repeated freeze-thaw cycles, and 100 μl of the supernatant was used for the second round of passage.

Genetic sequencing and sequence analysis. The deletion variants of SARS-CoV-2 were confirmed by the following approaches, as previously described (17): (i) 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 instrument; (ii) cDNA direct sequencing on an ONT MinION instrument; 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 targeted the 23444 to 23823 fragment of the viral genome (numbered according to GenBank accession number MN908947.3).

For metatranscriptomics, total RNAs were extracted from different types of samples by using the QIAamp viral RNA minikit, followed by DNase treatment and purification with Turbo DNase and Agencourt RNAClean XP beads. Libraries were prepared using the SMARTer stranded total transcriptome sequencing (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 PE150 platform. For the multiplex PCR approach, we followed the general method of multiplex PCR, as described in online (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 to 35 rounds of amplification, PCR products were collected, quantified, and then sequenced on an Illumina

MiSeq PE150 instrument or MinION sequencing device. Assembly of the Illumina raw data was performed using Geneious v11.0.3. 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 a depth of ≥20 as a threshold. For direct cDNA sequencing, we followed the Nanopore direct cDNA sequencing protocol (catalog number SQK-DCS109). Briefly, 100 ng viral RNA was reverse transcribed using the SuperScript IV first-strand synthesis system (Invitrogen, USA), followed by RNA chain digestion and second-strand synthesis. A total of 20-ng cDNA libraries were loaded onto a FLO-MIN106 flow cell. Generated sequences were mapped to the reference sequence of GenBank accession number MN908947.3 using minimap2. The maximum likelihood (ML) phylogeny for the 24 viral strain genomes was estimated with PhyML (21) using the HKY model (22) with 4 gamma rate categories for the substitution rate (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 was observed, the cell monolayers were scraped with the back of a pipette tip. The virus lysate was used for genetic sequencing and viral strain amplification. To assess the kinetics of virus replication, the titers of different viral strains were determined, and then the strains were inoculated into Vero-E6 and Vero cells at a multiplicity of infection (MOI) of 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 phosphate-buffered saline (PBS) to remove unattached virus. Cells were lysed at different times postinoculation, and total RNA was extracted by using RNeasy minikit (Qiagen, Germany). Cellular viral loads were calculated by using the SARS-CoV-2 reverse-transcription PCR (RT-PCR) kit (Daan Gene, Guangzhou, China), and the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene was quantified in parallel as an endogenous control.

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

ACKNOWLEDGMENTS

This work was supported by grants from Guangdong Provincial Novel Coronavirus Scientific and Technological Project (2020111107001) and Science and Technology Planning Project of Guangdong (2018B020207006). T.A.B. is supported by the Medical Research Council (MR/S007555/1). The Wellcome Centre for Human Genetics is supported by Wellcome Trust grant 203141/Z/16/Z. This work was supported by the European Research Council under the European Commission Seventh Framework Programme (FP7/2007-2013)/European Research Council grant agreement (614725-PATHPHYLODYN).

We are grateful for the helpful discussions from virological.org.

REFERENCES

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, Yuan M-L, Zhang Y-L, Dai F-H, Liu Y, Wang Q-M, Zheng J-J, Xu L, Holmes EC, Zhang Y-Z. 2020. A new coronavirus associated with human respiratory disease in China. Nature 580:E7. https://doi.org/10.1038/s41586-020-2202-3.
2. WHO. 2020. Coronavirus disease (COVID-2019) situation reports. WHO, Geneva, Switzerland.
3. Cui J, Li F, Shi Z-L. 2019. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 17:181–192. https://doi.org/10.1038/s41579-018-0118-9.
4. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181:281–292.e6. https://doi.org/10.1016/j.cell.2020.02.058.
5. Li F, Li W, Farzan M, Harrison SC. 2005. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science 309:1864–1868. https://doi.org/10.1126/science.1116480.
6. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, Li N, Guo Y, Li X, Shen X, Zhang Z, Shu F, Huang W, Li Y, Zhang Z, Chen R-A, Wu Y-J, Peng S-M, Huang M, Xie W-J, Cai Q-H, Hou F-H, Chen W, Xiao L, Shen Y. 2020. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature https://doi.org/10.1038/s41586-020-2313-x.
7. Lam TT-Y, Shum M-H, Zhu H-C, Tong Y-G, Ni X-B, Liao Y-S, Wei W, Cheung W-M, Li W-J, Li L-F, Leung GM, Holmes EC, Hu Y-L, Guan Y. 2020. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature https://doi.org/10.1038/s41586-020-2169-0.
8. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. Nat Med 26:450–452. https://doi.org/10.1038/s41591-020-0820-9.
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 Res 176:104742. https://doi.org/10.1016/j.antiviral.2020.104742.
10. Izaguirre G. 2019. The proteolytic regulation of virus cell entry by furin and other proprotein convertases. Viruses 11:837. https://doi.org/10.3390/v11090837.
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 https://doi.org/10.1101/2020.02.03.20019141.
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 https://doi.org/10.1101/2020.03.22.002204.

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. https://doi.org/10.1126/science.1092002.
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. Sci Rep 8:15177. https://doi.org/10.1038/s41598-018-33487-8.
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 181:271–274. https://doi.org/10.1016/j.cell.2020.02.052.
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 https://doi.org/10.1101/2020.03.02.974139.
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 181:997–1003.e9. https://doi.org/10.1016/j.cell.2020.04.023.
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, Alcantara LC, Jr, 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. Nat Protoc 12:1261–1276. https://doi.org/10.1038/nprot.2017.066.
19. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
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 Biol 20:8. https://doi.org/10.1186/s13059-018-1618-7.
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. https://doi.org/10.1093/sysbio/syq010.
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. https://doi.org/10.1007/BF02101694.
23. Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314. https://doi.org/10.1007/BF00160154.
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. https://doi.org/10.1016/j.gpb.2017.01.001.
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. https://doi.org/10.1093/nar/gkz913.