Natural deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape

Kevin R. McCarthy[1,2,3,†], Linda J. Rennick[1,2], Sham Nambulli[1,2], Lindsey R. Robinson-McCarthy[4], William G. Bain[5,6,7], Ghady Haidar[8,9], W. Paul Duprex[1,2,†]

Affiliations:

[1] Center for Vaccine Research, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[2] Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[3] Laboratory of Molecular Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
[4] Department of Genetics, Harvard Medical School, Boston, MA, USA
[5] Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Internal Medicine, UPMC, Pittsburgh, PA, USA
[6] Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[7] Staff Physician, VA Pittsburgh Healthcare System, Pittsburgh, PA, USA
[8] Division of Infectious Disease, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[9] Division of Infectious Disease, Department of Internal Medicine, UPMC, Pittsburgh, PA, USA

[†] Corresponding authors: Kevin R. McCarthy (krm@pitt.edu) and W. Paul Duprex (pduprex@pitt.edu)

Running title: SARS-CoV-2 spike evolves via deletion

Abstract:

Zoonotic pandemics follow the spillover of animal viruses into highly susceptible human populations. Often, pandemics wane, becoming endemic pathogens. Sustained circulation requires evasion of protective immunity elicited by previous infections. The emergence of SARS-CoV-2 has initiated a global pandemic. Since coronaviruses have a lower substitution rate than other RNA viruses this gave hope that spike glycoprotein is an antigenically stable vaccine target. However, we describe an evolutionary pattern of recurrent deletions at four antigenic sites in the spike glycoprotein. Deletions abolish binding of a reported neutralizing antibody. Circulating SARS-CoV-2 variants are continually exploring genetic and antigenic space via deletion in individual patients and at global scales. In viruses where substitutions are relatively infrequent, deletions represent a mechanism to drive rapid evolution, potentially promoting antigenic drift.

Main text:

SARS-CoV-2 emerged from a yet to be defined animal reservoir in 2019 and initiated a global pandemic (*1-5*). Currently there have been in excess of 56 million confirmed cases and 1.35 million recorded deaths (*6*). The scope of this pandemic suggests that SARS-CoV-2 will follow the trajectory of other emergent human respiratory viruses: a pandemic phase followed by establishment of an endemic human pathogen. The best-studied comparators come from influenza viruses, which have followed this course on four consecutive instances over the past century, and other coronaviruses, for example OC43 (*7*).

The transition from a pandemic to endemic pathogen is an evolutionary process. Endemic viruses evade immunity imparted by previous infection, typically by introducing substitutions in their glycoprotein(s) that disrupt the binding of protective antibodies. Influenza possesses an error-prone RNA-dependent RNA polymerase (RdRp), but often requires years to amass a sufficient subset of substitutions to alter antigenicity markedly (*8-10*). Coronavirus RdRps have proofreading activity,

and accordingly have much lower rates of nucleotide substitution than other RNA viruses (*11-13*). This slower rate of molecular evolution has provided hope that SARS-CoV-2 spike (S) glycoprotein will acquire limited antigenic diversity such that first-generation vaccines will provide durable immunity and protection from all circulating variants (*14, 15*).

We have identified an evolutionary signature defined by recurrent deletions at discrete sites within the S protein. This deletion mechanism rapidly introduces variation at antigenic sites of SARS-CoV-2. Deletions are observed in viruses sequenced from both chronically infected immunosuppressed patients and at a global scale. Deletions are most frequent at four sites, which we term recurrent deletion regions (RDRs). All four RDRs reside in the amino (N) terminal domain (NTD) of the S glycoprotein at defined antigenic sites. Deletions rapidly produce genetic novelty, including one variant that accounts for >2.5% of all sampled circulating viruses as of October 24, 2020. We have discovered that SARS-CoV-2 recurrently explores antigenic diversity, via deletion, producing variants that transmit between people. Importantly, deletions alter the epitope of a reported neutralizing antibody (*16*) and prevent its binding.

Natural deletions within the spike amino terminal domain arise independently during persistent human infections

An immunocompromised cancer patient infected with SARS-CoV-2 was treated in Pittsburgh. The patient was unable to clear the virus, despite treatment with Remdesivir and two infusions of convalescent serum. Significant amounts of virus were present in this individual when they ultimately succumbed to the infection 74 days after COVID-19 diagnosis (Hensley *et al*., submitted MS ID#: MEDRXIV/2020/234443). We consensus sequenced and cloned the S gene from these late time points directly from clinical material and identified two variants with deletions in the NTD (Fig. 1A). We term this individual Pittsburgh long-term infection 1 (PLTI1).

These data from PLTI1 prompted us to interrogate a number of patient metadata sequences deposited in GISAID (*17*). In searching for viruses similar to those obtained from PLTI1 we found eight patients with deletions in the S protein that had viruses sampled longitudinally over a period of weeks to months (Figs. 1A and S1A). For each, early time points had intact S sequences and at later time points deletions within the S gene. Six had deletions that were identical to, overlapping with or adjacent to those in PLTI1. Deletions at a second site developed in the other two patients (Fig. 1B). Viruses from seven patients possessed unique constellations of substitutions that were present at both early and late time points (Fig. S1B). These differentiate the viruses from each patient and strongly suggest that the deletion variants were not acquired in the community or nosocomially. Two unrelated patients with similar deletions have been recently reported by Avanzato & Matson (*18*) and Choi & Choudhary (*19*) and their respective colleagues. These sequences are included in our analysis. The most parsimonious explanation is that each deletion arose independently in response to a common and strong selective pressure, to produce strikingly convergent outcomes.

Recurrent and convergent deletions occur in the SARS-CoV-2 NTD

We searched the GISAID sequence database (*17*) for additional instances of deletions within S protein. From a dataset of 146,795 sequences (deposited from 12/01/2019 to 10/24/2020) we identified 1,108 viruses with deletions in the S gene. When mapped to the S gene, 90% occupied four discrete sites within the NTD (Fig. 2A). We term these sites recurrent deletion regions (RDRs) and number them 1-4 from the 5’ to 3’ end of the gene. RDR2 corresponds to the deletion in Fig. 1A and RDR4 to Fig. 1B.

The vast majority of deletions appear to have arisen and been subsequently retained in replicating viruses. In-frame deletions should occur one third of the time and are multiples of three nucleotides. We observed a preponderance of in-frame deletions with lengths of 3, 6, 9 and 12 (Fig. 2B). Among

all deletions, 93% are in frame and do not produce a stop codon (Fig. 2C). In the NTD, >97% of deletions maintain the open reading frame, with most mapping to RDRs 1 to 4. Other spike domains do not follow this trend. Deletions in the receptor binding domain (RBD) and S2 preserve the reading frame 30% and 37% of the time, respectively. Tolerance and enrichment for deletions are therefore an intrinsic feature of RDRs.

The RDRs harbor a spectrum of deletions, from those that appear only in a single virus to those that are frequent in length and position. Deletions at RDRs 1 and 3 were strongly biased to a single site while RDRs 2 and 4 are composed of many different overlapping deletions. Preferences to remove specific nucleotides are apparent from the histograms in Fig. 2D. For all four RDRs, it appears that selection and perhaps transmission favors specific deletions over others.

We compared the geographic distribution and GISAID clade designations of viruses with deletions in RDRs to our entire dataset (Fig. 2E-F). Viruses with deletions in RDRs 2 and 4 generally reflected the geographic and genetic diversity in the GISAID database. This patterning is consistent with recurrent, independent deletion events at these sites. In contrast, viruses with deletions at RDRs 1 and 3 were overwhelmingly from Europe (and Oceania for RDR3) and from clades G and GR respectively. This indicates that viruses recurrently explore deletions at RDRs 1 and 3, and selection has favored specific deletions, in certain clades that circulate in limited geographies.

SARS-CoV-2 RDR variants transmit naturally between humans

The geographic and genetic distributions of some RDR variants suggest human-to-human transmission. We identified, for each RDR, instances where viruses with identical deletions were isolated from different patients around the same time. Two patients in France (male, age 58, EPI_ISL_582112 and female, age 59, EPI_ISL_582120) were found to have viruses that where 100% identical, including a six-nucleotide deletion in RDR1. We identified a cluster of four

individuals in Senegal that shared a three-nucleotide deletion in RDR2 and a deletion in Orf1ab (1605 to1608). These viruses group together among all Senegalese samples (Fig. 3A). The RDR2 deletion is identical to those in PLTI1, MSK-4, MSK-6 and MSK-8, demonstrating that this mutation arises independently and transmits between humans. Four patients from Ireland had viruses that share a three-nucleotide deletion in RDR3. These sequences form distinct branches among Irish SARS-CoV-2 sequences (Fig. 3B). A cluster of sequences from Switzerland, from at minimum two individuals, share a nine nucleotide deletion in RDR4 (Fig. 3C).

These examples are illustrative. Most sequences lack sufficient accompanying data to distinguish between recurrent sampling of a single patient or viruses from multiple patients. We found 599 sequences with the same three-nucleotide deletion in RDR1 that were sequenced by centers across the United Kingdom (UK). Similarly, other sequences from the UK either shared three-nucleotide deletions in RDR2 (n=87) or RDR3 (n=48). We examined the prevalence of RDR variants throughout the global pandemic from December 2019 to October 2020 (Fig. 3D). Representatives at each site are present throughout. Deletions at RDRs 1 and 3 were the most frequent. For these, a single variant, Δ69-70 in RDR1 and Δ210 in RDR3, predominate (Fig. 3E). RDR2 deletions appear to be more diverse with Δ145 predominating. The Δ69-70 variant has rapidly increased in abundance, from 0.01% of all viral sequences in July 2020 to ~2.5% in October 2020 (1^{st} to 24^{th}). The frequencies of Δ69-70, Δ210, and likely Δ145 with a rise and fall pattern, are best explained by bursts of natural transmission between humans.

SARS-CoV-2 RDR variants abolish binding of a reported neutralizing antibody

The recurrence and convergence of RDR deletions, particularly during long-term infections, is indicative of selection and escape from a common and strong selective pressure. RDRs 2 and 4 and RDRs 1 and 3 occupy two distinct surfaces on the S protein NTD (Fig. 4A). Both sites are the targets of antibodies (*16, 20, 21*). The epitope for neutralizing antibody 4A8 is formed entirely by

beta sheets and their extended connecting loops that harbor RDRs 2 and 4. We generated a panel of S protein mutants representing the four RDRs. We transfected cells with plasmids expressing these mutants and used indirect immunofluorescence to determine if RDR deletions modulated 4A8 binding. The two RDR2 deletions and one RDR4 deletion completely abolished binding of 4A8 whilst still allowing recognition by a monoclonal antibody targeting the S protein RBD (Fig. 4B). Deletions at RDRs 1 and 3 had no impact on the binding of either monoclonal antibody, confirming that they alter independent sites. Convergent evolution operates both within single RDRs and between RDRs to produce functionally equivalent adaptions by altering the same epitope. These observations demonstrate that naturally arising and circulating variants of SARS-CoV-2 S have altered antigenicity.

Discussion:

Historically, pandemics have waned and left behind endemic human pathogens. This transition is contingent upon evading immunity imparted by previous infection. Influenza viruses exemplify this pattern, having followed it at least four successive times in the past century. Unlike the error-prone RdRps of most human respiratory pathogens, coronaviruses like SARS-CoV-2 possess polymerases with proofreading activity (*11-13*). However, proofreading cannot correct deletions, which can rapidly alter entire stretches of amino acids and the structures they form. We have identified an evolutionary signature defined by prevalent and recurrent deletions in the S protein. Deletion is followed by human-to-human transmission of variants with altered antigenicity. The simplicity of using deletion to drive diversity is biologically compelling.

COVID-19 typically resolves within weeks, before the full maturation of humoral immunity to SARS-CoV-2 (*22-25*). During pandemics neither the infected patient nor subsequently infected individuals impart an immunological pressure on the virus. However, during a long-term persistent infection, virus replicates in the presence of endogenous or supplemented (e.g. convalescent sera

or therapeutic monoclonal) antibody mediated immunity. Viral evolution in such patients may foreshadow preferred avenues of adaption in immune experienced populations. In individuals, multiple variants with distinct deletions can arise over time, essentially existing as an intra-host quasispecies (*18, 19*). Comparisons between deletions arising independently in persistently infected individuals show striking recurrent and convergent evolution. At global scales, similar variants sporadically arise in different geographies and viral lineages. From these data, it is evident that SARS-CoV-2 is continuously exploring sequence and antigenic space in different genetic, environmental and geographical contexts. These processes have produced at least once a variant that accounts for 2.5% of the viruses sequenced and deposited in databases as of October 2020.

In the three-dimensional structure of the S protein, RDRs occupy two distinct surfaces. Antibodies directed to each have been identified (*16, 20, 21*). Deletions in RDR2 or RDR4 abolish the binding of neutralizing antibody 4A8 and are representative of a pattern of recurrent, convergent, evolution within and between sites. Most humans are likely to mount 4A8-equivalent responses; indeed antibody 4-8, from an unrelated donor, engages an overlapping epitope (*21*). The propagation of recombinant vesicular stomatitis viruses bearing the S glycoprotein in the presence of immune sera selects for mutations in RDR2 that confer neutralization resistance to serum antibodies from multiple patients (*26*). The deletions in RDRs 1 and 3 occupy an epitope that was structurally defined using Fabs produced from the convalescent serum of patient COV57 (*20*). The recurrent selection for deletions *in vivo*, their correspondence with defined epitopes, and their impact on antibody binding demonstrate that RDR variants alter the antigenicity of S protein.

The most recent sequences in our dataset are strongly biased to the UK and we show many variants with deletions in RDRs 1, 2, and 3 circulated widely across England, Northern Ireland, Scotland and Wales. These deletions alter one antigenic site (*16, 21, 26*) and likely alter another. The UK is a site for at least one Phase III trial of a SARS-CoV-2 vaccine. Given that deletion

variants alter the antigenicity of SARS-CoV-2 S protein, potential mismatches between circulating and vaccine candidates may confound estimates of efficacy.

SARS-CoV-2 appears to be on a trajectory to become an endemic human pathogen and antigenic sites will continue evolving to evade preexisting immunity. Deletions that rapidly alter entire stretches of amino acids at specific antigenic sites are already playing an important role. Efforts to track and monitor these recurrent, rapidly arising, geographically widespread variants are vital.

References:

1. N. Zhu *et al.*, A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727-733 (2020).
2. F. Wu *et al.*, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).
3. H. Zhou *et al.*, A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr Biol* **30**, 3896 (2020).
4. T. T. Lam *et al.*, Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282-285 (2020).
5. M. F. Boni *et al.*, Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* **5**, 1408-1417 (2020).
6. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* **20**, 533-534 (2020).
7. L. Ren *et al.*, Genetic drift of human coronavirus OC43 spike gene during adaptive evolution. *Sci Rep* **5**, 11451 (2015).
8. J. M. Fonville *et al.*, Antibody landscapes after influenza virus infection or vaccination. *Science* **346**, 996-1000 (2014).

9. T. Bedford *et al.*, Integrating influenza antigenic dynamics with molecular evolution. *Elife* **3**, e01914 (2014).

10. D. J. Smith *et al.*, Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371-376 (2004).

11. S. Duffy, L. A. Shackelton, E. C. Holmes, Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**, 267-276 (2008).

12. M. R. Denison, R. L. Graham, E. F. Donaldson, L. D. Eckerle, R. S. Baric, Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol* **8**, 270-279 (2011).

13. E. Minskaia *et al.*, Discovery of an RNA virus 3'->5' exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc Natl Acad Sci U S A* **103**, 5108-5113 (2006).

14. B. Dearlove *et al.*, A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc Natl Acad Sci U S A* **117**, 23652-23662 (2020).

15. J. W. Rausch, A. A. Capoferri, M. G. Katusiime, S. C. Patro, M. F. Kearney, Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proc Natl Acad Sci U S A* **117**, 24614-24616 (2020).

16. X. Chi *et al.*, A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **369**, 650-655 (2020).

17. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, (2017).

18. V. A. Avanzato *et al.*, Case Study: Prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised cancer patient. *Cell*, (2020).

19. B. Choi *et al.*, Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N Engl J Med*, (2020).

20. C. O. Barnes *et al.*, Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. *Cell* **182**, 828-842 e816 (2020).

21. L. Liu *et al.*, Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450-456 (2020).

22. X. He *et al.*, Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* **26**, 672-675 (2020).

23. J. Bullard *et al.*, Predicting infectious SARS-CoV-2 from diagnostic samples. *Clin Infect Dis*, (2020).

24. R. Wolfel *et al.*, Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**, 465-469 (2020).

25. W. D. Liu *et al.*, Prolonged virus shedding even after seroconversion in a patient with COVID-19. *J Infect* **81**, 318-356 (2020).

26. Y. Weisblum *et al.*, Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife* **9**, (2020).

27. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066 (2002).

28. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).

29. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**, 1641-1650 (2009).

30. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

31. A. Watanabe *et al.*, Antibodies to a Conserved Influenza Head Interface Epitope Protect by an IgG Subtype-Dependent Mechanism. *Cell* **177**, 1124-1135 e1116 (2019).

32. M. A. Moody *et al.*, H3N2 influenza infection elicits more cross-reactive and less clonally expanded anti-hemagglutinin antibodies than influenza vaccination. *PLoS One* **6**, e25797 (2011).

33. K. H. D. Crawford *et al.*, Protocol and Reagents for Pseudotyping Lentiviral Particles with SARS-CoV-2 Spike Protein for Neutralization Assays. *Viruses* **12**, (2020).

34. W. B. Klimstra *et al.*, SARS-CoV-2 growth, furin-cleavage-site adaptation and neutralization using serum from acutely infected hospitalized COVID-19 patients. *J Gen Virol*, (2020).

Acknowledgments: We thank all of the researchers from around the world who have made SARS-CoV-2 sequences available for use in the GISAID database. We thank Stephen C. Harrison for his support. We thank Dr. Alison Morris, Dr. Bryan McVerry, Dr. Georgios Kitsios, Dr. Barbara Methe, Heather Michael, Michelle Busch, John Ries, and Caitlin Schaefer at the University of Pittsburgh as well as the physicians, nurses, and respiratory therapists at the University of Pittsburgh Medical Center Shadyside-Presbyterian Hospital intensive care units for assistance with collection and processing of the endotracheal aspirate sample

Competing interests: The authors declare no competing interests.

Funding: This work was supported by The University of Pittsburgh, the Center for Vaccine Research, The Richard King Mellon Foundation, the Hillman Family Foundation (WPD) and UPMC Immune Transplant and Therapy Center (WGB, GH),

Data availability: Sequences from PLTI1 were deposited in NCBI GenBank under accession numbers MW269404 and MW269555. All other sequences are available via the GISAID SARS-CoV-2 sequence database (gsaid.org).

Supporting information:

Material and Methods

Figure S1 – S2.

Table S1.

Supplementary References

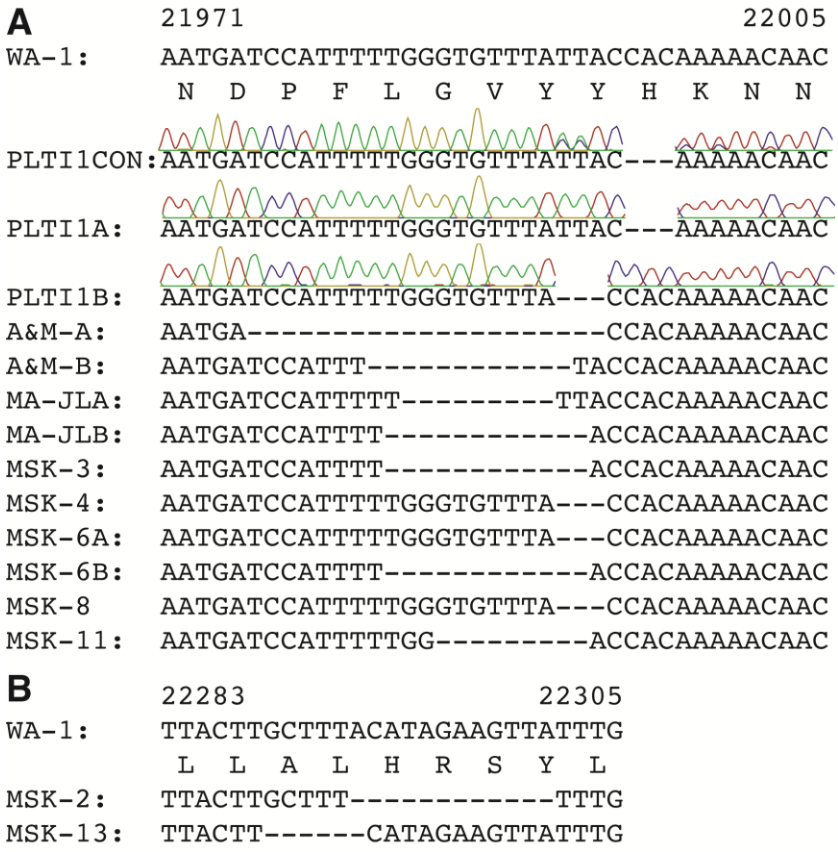

Figure 1. Deletions in SARS-CoV-2 spike arise during long-term persistent infections in immunosuppressed patients. A. Sequences of viruses isolated from PLTI1 and viruses from patients with deletions in the same NTD region. Chromatograms are shown for sequences from PLTI1, which include sequencing of bulk reverse transcription products and individual cDNA clones. Sequences from a patient reported by Avanzato & Matson and colleagues (*18*) are included and designated A&M and those reported from Choi & Choudhary and colleagues,(*19*) are designated MA-JL. Letters (A&B) designate different variants from the same patient. (B) Sequences of viruses from two patients with deletions in a different regions of the NTD. All sequences are aligned to the WA-1 reference sequence (MN985325).

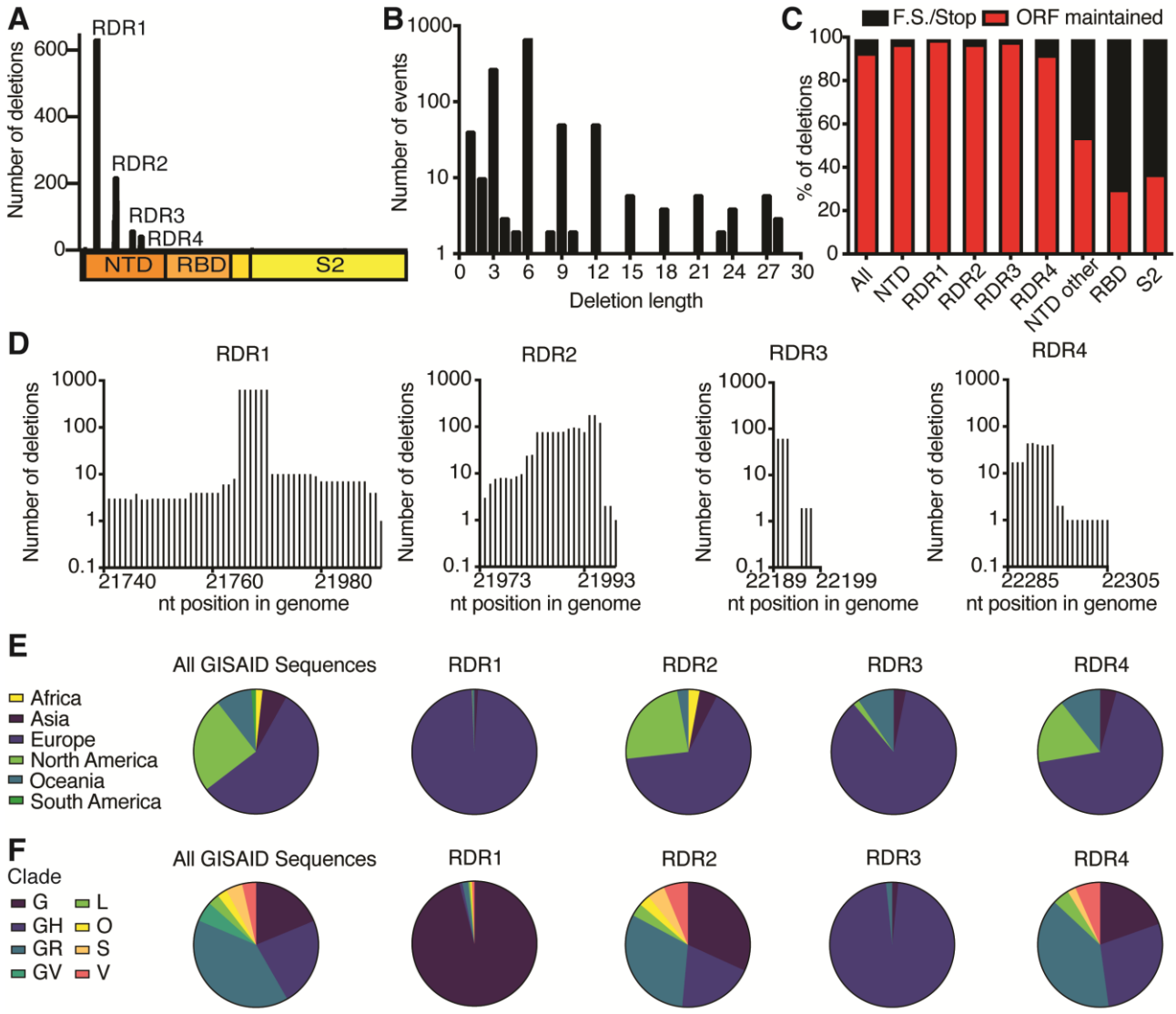

Figure 2. Identification and characterization of recurrent deletion regions in SARS-CoV-2 spike protein. A. The number of deletion events identified among sequences in the GISAID SARS-CoV-2 sequence database mapped to the S gene. These form four clusters, termed recurrent deletion regions (RDRs). B. Length distribution of deletion events shows a preference for deletions that preserve the reading frame. C. The percentage of deletion events at the indicated site that either maintain the open reading frame or introduce a frameshift or premature stop codon (F.S./Stop). D. Abundance of nucleotide deletions in each RDR. Positions are defined by reference sequence MN985325. E and F. Geographic (E) and genetic (F) distributions of RDR variants compared to the entire GISAID database (sequences from 12-1-2019 to 10-24-2020). GISAID clade classifications are used in F.

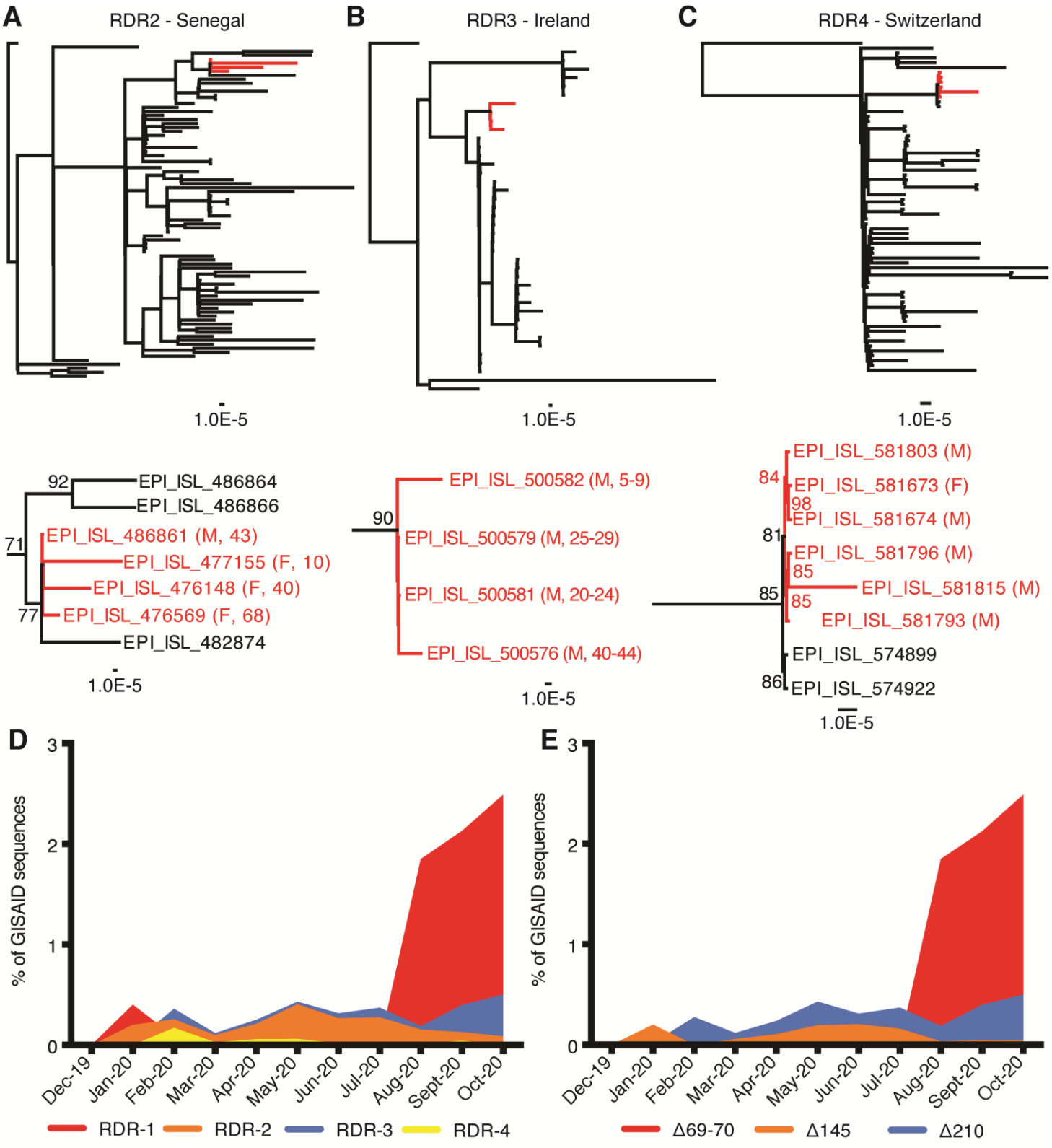

Figure 3. SARS-CoV-2 viruses with spike deletions transmit between humans. Maximum likelihood phylogenetic trees are rooted on MN985325 and were calculated with 10,000 (A and B) or 1000 (C) bootstrap replicates. Branches with transmitted RDR variants are colored red and detailed

below. Patient data differentiating individual patients is provided. For clarity, bootstrap values below 70 are removed. Supporting figure S2 provides all branch labels. A. Transmission of an RDR2 variant among 4 individuals in Senegal (deletion positions 21,991-21,994). B. Transmission cluster of an RDR3 variant (deletion positions 22,189-22,192) among four Irish patients. C. Transmission of an RDR4 variant (deletion positions 22,281-22,290) among at least one male and female in Switzerland. D. Frequency of RDR variants among all complete genomes deposited in GISAID between December 2019 and October 24, 2020. E. Frequency of specific RDR deletion variants (numbered according to spike amino acids) among all GISAID variants over the same time period. The plot of RDR3/Δ210 has been adjusted by 0.02 units on the Y-axis for visualization in panel D due to its overlap with RDR2 and this adjustment has been retained in panel E to make direct comparisons between panels.

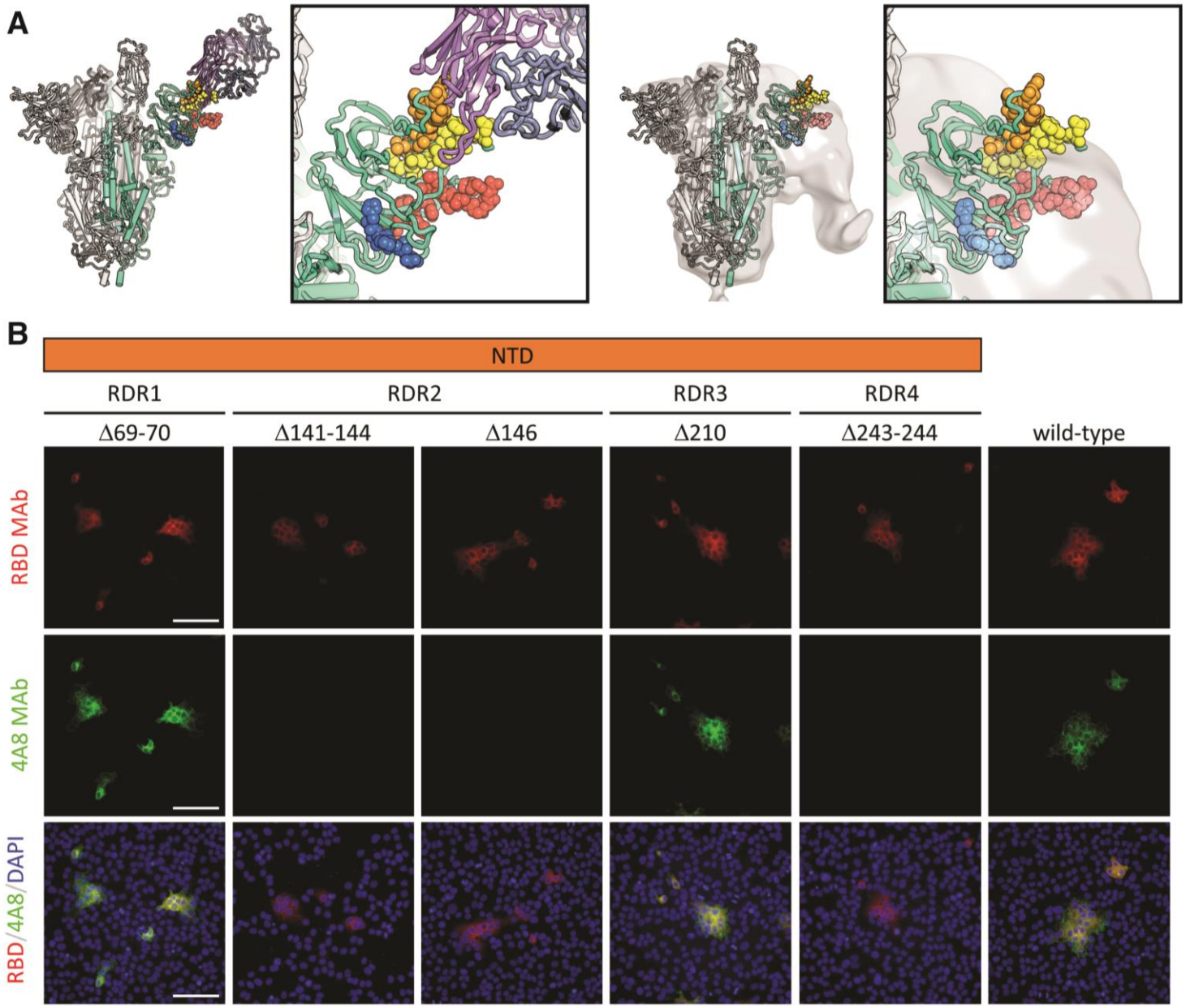

Figure 4. Deletions in the spike NTD alter its antigenicity. RDRs map to defined antigenic sites. A. Top: A structure of antibody 4A8 (*16*) (PDB: 7C21) (purples) bound to one protomer (green) of a SARS-CoV-2 spike trimer (grays). RDRs 1-4 are colored red, orange, blue, and yellow, respectively, and shown in spheres. The interaction site is shown at right. Bottom: The electron microscopy density of COV57 serum Fabs (*20*) (EMDB emd_22125) fit to SARS-CoV-2 Spike trimer (PDB: 7C21). The same view of the interaction site is provided at right. B. S protein distribution in Vero E6 cells at 24 h post-transfection with S protein deletion mutants, visualized by immunodetection in permeabilized cells. A monoclonal antibody against SARS-CoV-2 S protein receptor-binding domain

(RBD MAb; red) detects all mutant forms of the protein (Δ69-70, Δ141-144, Δ146, Δ210, Δ243-244) and the unmodified protein (wild-type). 4A8 monoclonal antibody (4A8 MAb; green) does not detect mutants containing deletions in RDR2 or RDR4 (Δ141-144, Δ243-244, Δ146). Overlay images (RBD/4A8/DAPI) depict co-localization of the antibodies; nuclei were counterstained with DAPI (blue). The scale bars represent 100 μm.

Supporting information:

Methods:

Determination of PLTI1 patient spike gene sequences: To determine the consensus sequence of SARS-CoV-2 S in the patient endotracheal aspirate sample collected at day 72 (Hensley *et al*., submitted MS ID#: MEDRXIV/2020/234443), RNA was isolated from the sample using TRIzol LS (Thermo Fisher Scientific), cDNA was generated using the Superscript III first strand synthesis system (Thermo Fisher Scientific) and random hexamers, DNA was amplified using Phusion DNA polymerase (New England BioLabs) and SARS-CoV-2 specific primers surrounding the open reading frame for the spike protein, and the consensus sequence was determined by Sanger sequencing (Genewiz) using SARS-CoV-2 specific primers. The amplified DNA product was also cloned into pCR Blunt II TOPO vector using a Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher Scientific) and the spike NTD sequence of individual clones was determined by Sanger sequencing (Genewiz) using M13F and M13R primers. Individual clone sequences are available with accession numbers MW269404 and MW269555.

Sequence analysis: Sequences were obtained from the publically available GISAID database and acknowledged in supporting Table 1. Our dataset was composed of SARS-CoV-2 sequences collected and deposited between 12-1-19 and 10-24-20. Sequence analysis was performed in Geneious (Biomatters, New Zealand). To identify deletion variants in S gene, sequences were mapped to NCBI reference sequence MN985325 (SARS-CoV-2/human/USA/WA-CDC-WA1/2020), the S gene open reading frame was extracted, remapped to reference and parsed for deletions using a search for gaps function. Sequences with deletions were manually extracted for subsequent analysis.

All identified deletion variants and MN985325 were aligned using MAFFT (*27, 28*) and adjusted manually in recurrent deletion regions for consistency.

Phylogenetic analyses utilized all sequences in our dataset from a country at a specific time, or in the case of Senegalese sequences the entirety of the pandemic. For non-Senegalese samples, sequences obtained within 1-2 months of the variants of interest were aligned to MN985325 using MAFFT (*27, 28*). FastTree (*29*) was used to generate a preliminary phylogeny from which we extracted the sequences corresponding to the lineage of interest and adjacent outgroups. These sequences were realigned using MAFFT. Maximum- Likelihood phylogenetic trees were calculated using RAxML (*30*) using a general time reversible model with optimization of substitution rates (GTR GAMMA setting), starting with a completely random tree, using rapid Bootstrapping and search for best-scoring ML tree. Between 1,000 and 10,000 bootstraps of support were performed.

Cell lines: Human 293F cells were maintained at 37° Celsius with 5% CO_2 in FreeStyle 293 Expression Medium (ThermoFisher) supplemented with penicillin and streptomycin. Vero E6 cells were maintained at 37° Celsius with 5% CO_2 in high glucose DMEM (Invitrogen) supplemented with 1% (v/v) Glutamax (Invitrogen) and 10% (v/v) fetal bovine serum (Invitrogen).

Recombinant IgG expression and purification: The heavy and light chain variable domains of 4A8 (*16*) was synthesized by Integrated DNA Technologies (Coralville, Iowa) and cloned into a modified human pVRC8400 expression vector encoding for full length human IgG1 heavy chains and human kappa light chains. Plasmids encoding influenza hemagglutinin-specific antibody H2214 have been described previously (*31, 32*). IgGs were produced by polyethylenimine (PEI) facilitated, transient transfection of 293F cells that were maintained in FreeStyle 293 Expression Medium. Transfection complexes were prepared in Opti-MEM and added to cells. Five days post-transfection (d.p.t.) supernatants were harvested, clarified by low-speed centrifugation, adjusted to pH 5 by

addition of 1 M 2-(N-morpholino)ethanesulfonic acid (MES) (pH 5.0), and incubated overnight with Pierce Protein G Agarose resin (Pierce, ThermoFisher). The resin was collected in a chromatography column, washed with a column volume of 100 mM sodium chloride 20 mM (MES) (pH 5.0) and eluted in 0.1 M glycine (pH 2.5) which was immediately neutralized by 1 M TRIS(hydroxymethyl)aminomethane (pH 8). IgGs were then dialyzed against phosphate buffered saline (PBS) pH 7.4.

Cloning and transfection of SARS-CoV-2 spike protein deletion mutants: A series of deletion mutants were generated in HDM_SARS2_Spike_del21_D614G (*33*) a plasmid containing SARS-CoV-2 S protein lacking the 21 C-terminal amino acids. HDM_SARS2_Spike_del21_D614G was a gift from Jesse Bloom (Addgene plasmid # 158762; http://n2t.net/addgene:158762; RRID:Addgene_158762). Cloning strategies were designed to delete S protein amino acids 69-70 (Δ69-70), 141-144 (Δ141-144), 146 (Δ146), 210 (Δ210) or 243-244 (Δ243-244). Appropriate gBlocks were generated synthetically (Integrated DNA Technologies) and cloned into HDM_SARS2_Spike_del21_D614G by Gibson Assembly using NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs). Assemblies were transformed into DH5-alpha chemically competent cells (New England Biolabs) and correct clones were identified by restriction profile and Sanger sequencing (Genewiz) of small scale plasmid preparations from individual bacterial clones. Plasmid DNA for transfections was prepared using a HiSpeed Plasmid Midi Kit (Qiagen). Vero E6 cells were seeded into 24 well trays at 10^5 cells per well. After overnight incubation at 37° Celsius, 5% (v/v) CO_2, the cells were rinsed with Opti-MEM (Invitrogen), 1ml/well Opti-MEM was added and cells were incubated at 37° Celsius, 5% (v/v) CO_2 for 30 minutes. Transfection mixes were prepared, according to manufacturer's instructions, containing 200 ng/well of plasmid DNA with 3 µl per µg DNA of Lipofectamine 2000 (Invitrogen). After the 30 minute incubation Opti-MEM in the wells was replaced with 500 µl per well Opti-MEM and 100 µl per well of transfection mixes were added. Transfected cells were incubated at 37° Celsius, 5% (v/v) CO_2 for 24 hours.

Indirect immunofluorescence assay: Indirect immunofluorescence was performed as previously reported (*34*). Briefly, cells transfected with the SARS-CoV-2 S protein deletion mutants and controls were washed once with DPBS (Fisher Scientific), fixed with 4% (w/v) paraformaldehyde in PBS (Boston Bioproducts) for 20 minutes at room temperature, rinsed twice with DPBS and permeabilized with 0.1% (v/v) Triton-X100 (Sigma) in DPBS for 30 minutes at 37° Celsius. Primary antibodies [rabbit anti-SARS-CoV-2 S monoclonal antibody, 40150-R007, Sino Biological, 1/700 dilution and human 4A8 monoclonal antibody, 1 µg/ml, in PBS containing 0.1% (v/v) Triton X-100] were added and incubated at 37° Celsius for 1 hour. Cells were washed three times with DPBS and secondary antibodies [goat anti-rabbit Alexa Fluor-568, Invitrogen, and goat anti-human Alexa Fluor-488, Invitrogen, diluted 1:400 in DPBS containing 0.1% (v/v) Triton X-100 were added and incubated at 37° Celsius for 1 hour. Cells were washed three times with DPBS and nuclei were counterstained with 4',6-diamidino-2-phenylindole (DAPI) nuclear stain (300 nM DAPI stain solution in PBS; Invitrogen) for 10 minutes at room temperature. Fluorescence was observed with a DMi 8 UV microscope (Leica) and photomicrographs were acquired using a camera (Leica) and LAS X software (Leica). Appropriate controls were included to determine antibody specificity.

Structure visualization: Structural figures were rendered in Pymol (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC).

Supplementary Figures:

A

MA-JL-D		NY-MSK-2		NY-MSK-3		NY-MSK-4	
Date	**Accession**	**Date**	**Accession**	**Date**	**Accession**	**Date**	**Accession**
4/28/20	EPI_ISL_593478	3/30/20	EPI_ISL_583428	3/19/20	EPI_ISL_583431	4/13/20	EPI_ISL_583439
5/5/20	EPI_ISL_593479	3/30/20	EPI_ISL_583429	3/19/20	EPI_ISL_583434	4/13/20	EPI_ISL_583441
6/24/20	EPI_ISL_593480	4/23/20	EPI_ISL_583430	4/2/20	EPI_ISL_583432	5/6/20	EPI_ISL_583440
6/30/20	EPI_ISL_593553			4/2/20	EPI_ISL_583435	5/6/20	EPI_ISL_583442
8/16/20	EPI_ISL_593554			4/13/20	EPI_ISL_583433	5/28/20	EPI_ISL_583443
8/18/20	EPI_ISL_593555			4/13/20	EPI_ISL_583436		
8/31/20	EPI_ISL_593556			4/21/20	EPI_ISL_583437		
9/3/20	EPI_ISL_593557			5/11/20	EPI_ISL_583438		
9/9/20	EPI_ISL_593558						

NY-MSK-6		NY-MSK-8		NY-MSK-11		NY-MSK-13	
Date	**Accession**	**Date**	**Accession**	**Date**	**Accession**	**Date**	**Accession**
4/30/20	EPI_ISL_583446	4/20/20	EPI_ISL_583456	3/24/20	EPI_ISL_583466	3/17/20	EPI_ISL_583469
5/6/20	EPI_ISL_583447	4/20/20	EPI_ISL_583458	3/24/20	EPI_ISL_583467	5/4/20	EPI_ISL_583470
5/12/20	EPI_ISL_583448	5/7/20	EPI_ISL_583459	4/16/20	EPI_ISL_583468	7/18/20	EPI_ISL_583471
5/19/20	EPI_ISL_583449	5/7/20	EPI_ISL_583457				
5/25/20	EPI_ISL_583450	5/12/20	EPI_ISL_583460				
6/4/20	EPI_ISL_583451						
7/22/20	EPI_ISL_583452						
7/22/20	EPI_ISL_583453						
7/22/20	EPI_ISL_583454						

B

Patient identifier	Substitution 1	Substitution 2	Substitution 3	Substitution 4	Substitution 5	Substitution 6	Substitution 7
MA-JL-D	U1336C	G7936C	C16580T	C27881T	-	-	-
NY-MSK-2	C5392T	C27236T	-	-	-	-	-
NY-MSK-3	A2018G	G4790A	C24518T	-	-	-	-
NY-MSK-4	A17376G	G28881A	G28882A	G28883C	-	-	-
NY-MSK-6	G6358A	G11083T	C11916U	C18998U	U25233C	C25603T	G29540A
NY-MSK-8	-	-	-	-	-	-	-
NY-MSK-11	C19593U	-	-	-	-	-	-
NY-MSK-13	C920U	-	-	-	-	-	-

Figure S1. Information for the longitudinally sampled patients that were identified in the GISAID database and detailed in Figure 1. (A) Date of collection and GISAID accession number for each sequence. (B) Identifying substitutions unique to each patient among the longitudinally sampled patients reported here.

Fig. S2

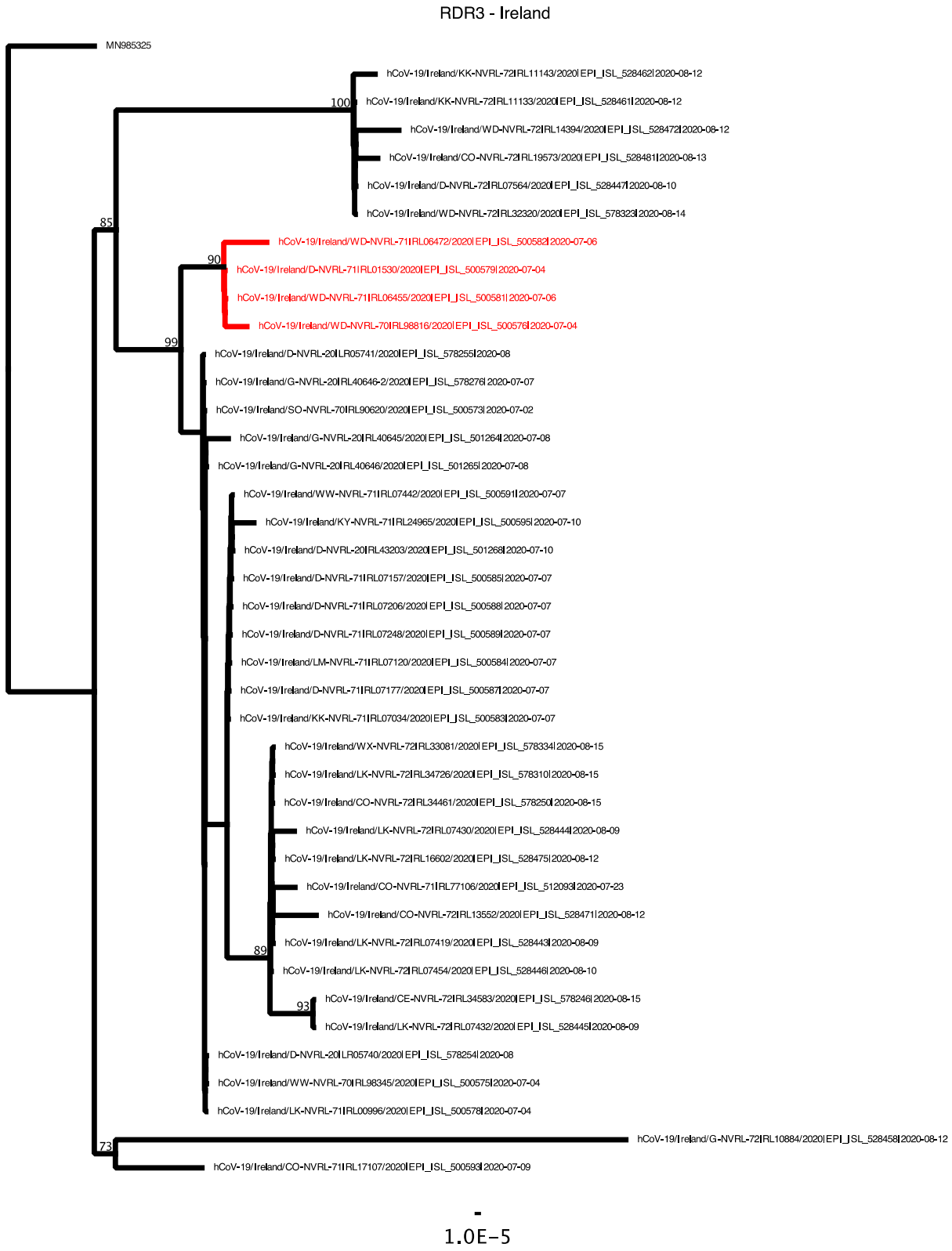

Fig. S2 continued

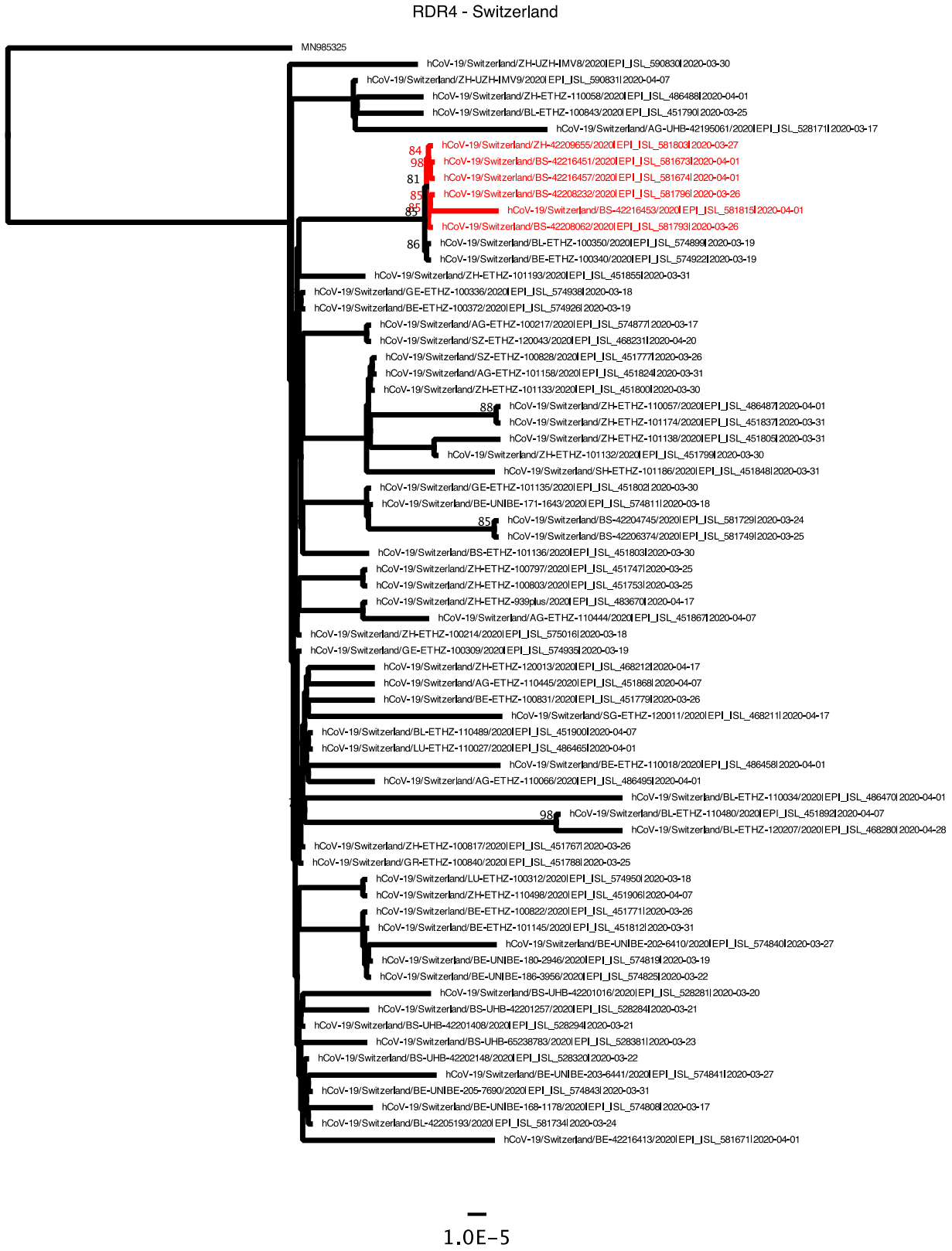

Figure S2. Phylogenetic analysis of transmitted RDR variants. The trees from Figure 3 are shown with branch labels. For clarity nodes with bootstrap values above 70 are labeled.

Supplemental References:

1. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066 (2002).
2. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).
3. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**, 1641-1650 (2009).
4. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
5. X. Chi *et al.*, A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **369**, 650-655 (2020).
6. A. Watanabe *et al.*, Antibodies to a Conserved Influenza Head Interface Epitope Protect by an IgG Subtype-Dependent Mechanism. *Cell* **177**, 1124-1135 e1116 (2019).
7. M. A. Moody *et al.*, H3N2 influenza infection elicits more cross-reactive and less clonally expanded anti-hemagglutinin antibodies than influenza vaccination. *PLoS One* **6**, e25797 (2011).
8. K. H. D. Crawford *et al.*, Protocol and Reagents for Pseudotyping Lentiviral Particles with SARS-CoV-2 Spike Protein for Neutralization Assays. *Viruses* **12**, (2020).
9. W. B. Klimstra *et al.*, SARS-CoV-2 growth, furin-cleavage-site adaptation and neutralization using serum from acutely infected hospitalized COVID-19 patients. *J Gen Virol*, (2020).