Recurrent independent emergence and transmission of SARS-CoV-2 Spike amino acid H69/V70 deletions

Ravindra Gupta (rkg20@cam.ac.uk)
University of Cambridge
Steven Kemp
UCL
William Harvey
University of Glasgow
Spyros Lytras
University of Glasgow https://orcid.org/0000-0003-4202-6682
Alessandro Carabelli
University of Cambridge
David Robertson
Univeristy of Glasgow https://orcid.org/0000-0001-6338-0221

Biological Sciences - Article

Keywords: SARS-CoV-2, COVID-19, antibody escape, neutralising antibodies, mutation, evasion, resistance, fitness, evolution

Posted Date: January 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-136937/v1

Abstract

SARS-CoV-2 Spike amino acid replacements in the receptor binding domain (RBD) occur relatively frequently and some have a consequence for immune recognition. Here we report recurrent emergence and significant onward transmission of a six-nucleotide deletion in the S gene, which results in loss of two amino acids: H69 and V70. Of particular note this deletion, ΔH69/V70, often co-occurs with the receptor binding motif amino acid replacements N501Y, N439K and Y453F. One of the ΔH69/V70+ N501Y lineages, B.1.1.7, is comprised of over 4000 SARS-CoV-2 genome sequences from the UK and includes eight other S gene mutations: RBD (N501Y and A570D), S1 (ΔH69/V70 and Δ144/145) and S2 (P681H, T716I, S982A and D1118H). Some of these mutations have presumably arisen as a result of the virus evolving from immune selection pressure in infected individuals and at least one, lineage B.1.1.7, potentially from a chronic infection. Given our recent evidence that ΔH69/V70 enhances viral infectivity (Kemp et al. 2020), its effect on virus fitness appears to be independent to the RBD changes. Enhanced surveillance for the ΔH69/V70 deletion with and without RBD mutations should be considered as a priority. Permissive mutations such as ΔH69/V70 have the potential to enhance the ability of SARS-CoV-2 to generate new variants, including vaccine escape variants, that would have otherwise significantly reduced viral fitness.

Background

SARS-CoV-2's Spike surface glycoprotein engagement of hACE2 is essential for virus entry and infection[1], and the receptor is found in respiratory and gastrointestinal tracts[2]. Despite this critical interaction and the constraints it imposes, it appears the RBD, and particularly the receptor binding motif (RBM), is relatively toleratant to mutations[3,4], raising the real possibility of virus escape from vaccine-induced immunity and monoclonal antibody treatments. Spike mutants exhibiting reduced susceptibility to monoclonal antibodies have been identified in *in vitro* screens[3,5,6], and some of these mutations have been found in clinical isolates[7]. Due to the susceptibility of the human population to this virus, the acute nature of infections and limited use of vaccines to date there has been limited selection pressure placed SARS-CoV-2[8]; as a consequence few mutations that could alter antigenicity have increased significantly in frequency.

The unprecedented scale of whole genome SARS-CoV-2 sequencing has enabled identification and epidemiological analysis of transmission and surveillance, particularly in the UK[9]. As of December 18^{th,} there were 270,000 SARS-CoV-2 sequences available in the GISAID Initiative (https:gisaid.org/). However, geographic coverage is very uneven with some countries sequencing at higher rates than others. This could result in novel variants with altered biological or antigenic properties evolving and not being detected until they are already at high frequency.

Studying SARS-CoV-2 chronic infections can give insight into virus evolution that would require many chains of acute transmission to generate. This is because the majority of infections arise as a result of early transmission during pre or asymptomatic phases prior to peak adaptive responses, and virus

adaptation not observed as the virus is usually cleared by the immune response[10,11]. We recently documented *de novo* emergence of antibody evasion mutations mediated by S gene mutations in an individual treated with convalescent plasma (CP)[12]. Dramatic changes in the prevalence of Spike variants ΔH69/V70 (an out of frame six-nucleotide deletion) and D796H variant followed repeated use of CP, while *in vitro* the mutated ΔH69/V70 + D796H variant displayed reduced susceptibility to CP, at the same time retaining infectivity comparable to wild type[12]. The ΔH69/V70 itself conferred a two-fold increase in Spike mediated infectivity using pseudotyped lentiviruses. Worryingly, other deletions in the N-Terminal Domain (NTD) have been reported to arise in chronic infections[7] and provide escape from NTD-specific neutralising antibodies[13].

Here we analysed the available GISAID Initiative data for circulating SARS-CoV-2 Spike sequences containing ΔH69/V70 and performed phylogenetic and structural modelling from the pandemic data and across sarbecoviruses in different species. We find, while occurring independently, the Spike ΔH69/V70 often emerges after a significant RBM amino acid replacement that increases binding affinity to hACE2. Protein structure modelling indicates this mutation could also contribute to antibody evasion as suggested for other NTD deletions[13].

Results

The deletion H69/V70 is present in over 6000 sequences worldwide, 2.5% of the available data (Fig. 1, Supplementary Fig. 1), and largely in Europe from where most of the sequences in GISAID are derived (Fig. 1C-D). Many of the sequences are from the UK and Denmark where sequencing rates are high compared to other countries. ΔH69/V70 occurs in variants observed in different global lineages, representing multiple independent acquisitions of this SARS-CoV-2 deletion (Fig. 1A). While variants with deletions in this region of Spike are observed in GISAID[13], the earliest unambiguous sequence that includes the ΔH69/V70 was detected in Sweden in April 2020, an independent deletion event relative to other ΔH69/V70 variants. The prevalence of ΔH69/V70 has since increased since August 2020 (Fig. 1C-D). Further analysis of sequences revealed, firstly, that single deletions of either 69 or 70 were uncommon and, secondly, some lineages of ΔH69/V70 alone were present (Fig. 1A), as well as ΔH69/V70 in the context of other mutations in Spike, specifically those in the RBM (Fig. 1A, E, F).

To guage the importance of this part of Spike molecule, we examined the 69/70 region of Spike in a set of other known *Sarbecoviruses*, relatives to SARS-CoV-2 (Fig. 2). We observe substantial variability in the region, specifically caused by indels, with some viruses including SARS-CoV having 6–7 amino acid deletions (Fig. 2B). This is indicative of plasticity in this protein region that could allow the sarbecoviruses to alter their Spike conformation. The second closest relative to SARS-CoV-2 for this region after RaTG13 is the cluster of 5 CoVs sampled in trafficked pangolins in the Guangxi province[14]. Looking at the 69/70 region in these virus sequences raises the interesting observation that one of the five viruses in the cluster, P1E, has amino acids 69H and 70L present, while the other four have a double amino acid deletion (Fig. 2B). Given that SARS-CoV-2 and RaTG13 have the homologous HV insertion at

these positions, the most parsimonious explanation is that the proximal common ancestor between SARS-CoV-2 and the Guangxi pangolin cluster had the insertion, which was then lost while circulating in the pangolin population, similar to what we now see with SARS-CoV-2 in humans. Interestingly, the double amino acid deletion in the pangolin viruses in in-frame in contrast to what is seen in SARS-CoV-2 (e.g. lineage B.1.1.7, Fig. 2C).

To estimate the structural impact of ΔH69/V70, the protein structure of the NTD possessing the double deletion was modelled. The ΔH69/V70 deletion was predicted to alter the conformation of a protruding loop comprising residues 69 to 76, pulling it in towards the NTD (Fig. 3A). In the post-deletion structural model, the positions of the alpha carbons of residues either side of the deleted residues, Ile68 and Ser71, were each predicted to occupy positions 2.9 Å from the positions of His69 and Val70 in the pre-deletion structure. Concurrently, the positions of Ser71, Gly72, Thr73, Asn74 and Gly75 are predicted to have changed by 6.5 Å, 6.7 Å, 6.0 Å, 6.2 Å and 8 Å, respectively, with the overall effect of these residues moving inwards, resulting in a less dramatically protruding loop. The position of this loop in the structure prior to the occurrence of the ΔH69/V70 is shown in the context of the wider NTD in Fig. 3B. The locations of main RBD mutations observed with ΔH69/V70 are shown in Fig. 3C and D. Residues belonging to a similarly exposed, nearby loop that form the epitope of a neutralising, NTD-binding epitope are also highlighted.

We next examined the lineages where S gene mutations in the RBD were identified at high frequency, in particular co-occurring with N439K (Fig. 3C,D), an amino acid replacement reported to be defining variants increasing in numbers in Europe and other regions[3] (Fig. 1E, Supplementary Fig. 2). N439K appears to have reduced susceptibility to a small subset of monoclonals targeting the RBD, whilst retaining affinity for ACE2 *in vitro*[3]. The proportion of viruses with ΔH69/V70 only increased from August 2020 when it appeared with the second N439K lineage, B.1.141[3] (Fig. 1E). As of November 26^{th,} remarkably there were twice as many cumulative sequences with the deletion as compared to the single N439K indicating it may be contributing to the success of this lineage (Fig. 1E). Due to their high sampling rates the country with the highest proportion of N439K+ΔH69/V70 versus N439K alone is England. The low levels of sequencing in most countries indicate N439K's prevalence could be relatively high[3]. In Scotland, where early growth of N439K was high (forming N439K lineage B.1.258 that subsequently went extinct with other lineages after the lockdown[3]), there is now an inverse relationship with 546 versus 177 sequences for N439K and N439K+ΔH69/ΔV70 respectively (Fig. 1E). These differences therefore likely reflect differing epidemic growth characteristics and timings of the introductions the N439K variants with or without the deletion.

The second significant cluster with ΔH69/V70 and RBD mutants involves Y453F, another Spike RBD mutation that increases binding affinity to ACE2[4] (Fig. 3C,D) and has been found to be associated with mink-human infections[15]. In one SARS-CoV-2 mink-human sub-lineage, termed 'Cluster 5', Y453F and ΔH69/V70 occurred with F486L, N501T and M1229I and was shown to have reduced susceptibility to sera from recovered COVID-19 patients (https://files.ssi.dk/Mink-cluster-5-short-report_AFO2). Y453F has

been described as an escape mutation for mAb REGN10933[16]. The ΔH69/V70 was first detected in the Y453F background on August 24th and thus far appears limited to Danish sequences (Supplementary Fig. 3).

A third lineage containing the same deletion ΔH69/V70 has arisen with another RBD mutation N501Y (Fig. 4A, C, Supplementary Fig. 4). Based on its location it might be expected to escape antibodies similar to COV2-2499[5] (Fig. 3C, D). In addition, when SARS-CoV-2 was passaged in mice for adaptation purposes for testing vaccine efficacy, N501Y emerged and increased pathogenicity[17]. Sequences with N501Y alone were isolated both in the UK, Brazil and USA in April 2020, and recently in South Africa[18]. A newly described N501Y-derived lineage in South Africa, (B.1.351, Fig. 1A) (also termed 501Y.V2) is characterised by eight mutations in the Spike protein, including N501Y and two other important residues in the receptor-binding domain (K417N and E484K) which are important residues in RBM[18]. The positions of residues 417, 484, and 501 proximal to the bound hACE2 are shown in Fig. 4C. The E484K substitution has been identified as antigenically important being reported as an escape mutation for several monoclonal antibodies including C121, C144[19], REGN10933 and REGN10934[16]. The increase in hACE2 binding affinity caused by N501Y is permissive of the mutation K417N, specifically, ACE2 affinity induced by N501Y (+ 0.24 $\Delta \log_{10} K_d$) may be compensated by K417N (-0.45 $\Delta \log_{10} K_d$)[20]. Residue K417 is also identified as antigenically significant with K417E facilitating escape from mAb REGN10933[16].

N501Y + ΔH69/V70 sequences were first detected in the UK on 20th September 2020, with the cumulative number of N501Y + ΔH69/V70 mutated sequences now exceeding the single mutant N501Y lineage (Fig. 1F). On closer inspection these sequences were part of a new lineage (B.1.1.7), termed VOC 202012/01 by Public Health England as they are associated with relatively high numbers of infections (Fig. 4A-C, Supplementary Fig. 4). In addition to RBD N501Y + NTD ΔH69/V70 this new variant had five further S mutations across the RBD (N501Y) and S2 (P681H, T716I, S982A and D1118H), as well as NTD Δ144[21] (Fig. 4B). The variant has now been identified in a number of other countries, including Hong Kong, Japan, Australia, France, Spain, Singapore, Israel, Switzerland, and Italy. This lineage has a relatively long branch due to 23 unique mutations (Fig. 4A and supplementary Fig. 4), consistent within host evolution and spread from a chronically infected individual[21]. Notably a sequence can be identified that contains ΔH69/V70, N501Y and D1118H (Fig. 4A, black box). However, the available sequences did not enable us to determine whether the B1.1.7 mutations N501Y + ΔH69/V70 arose as a result of a N501Y virus acquiring ΔH69/V70 or vice versa.

The B.1.1.7 lineage has some notable features. Firstly the Spike Δ144 mutation could lead to loss of binding of the S1-binding neutralising antibody 4A8[13] (Fig. 3B). The Y144 sidechain is itself around 4.5 Å from the nearest atoms of 4A8 complexed with Spike and the deletion is expected to alter the positions neighbouring residues that directly interact with 4A8; contacting residues (145, 146, 147, 150, 152, 246 and 258)[22] are shown in magenta in Fig. 3B. Secondly the P681H mutation lies within the furin cleavage site. Furin cleavage is a property of some more distantly related coronaviruses, and in particular not found in SARS-CoV-1[23]. When SARS-CoV-2 is passaged *in vitro* it results in mutations in the furin

cleavage site, suggesting the cleavage is dispensable for *in vitro* infection[24]. The significance of furin site mutations may be related to potential escape from the innate immune antiviral IFITM proteins by allowing infection independent endosomes[25]. The significance of the multiple S2 mutations is unclear at present, though D614G, also in S2 was found to lead to a more open RBD orientation to explain its higher infectivity[26]. T716I and D1118H occur at residues located close to the base of the ectodomain (Fig. 4B) that are partially exposed and buried, respectively. The residue 982 is located centrally, in between the NTDs, at the top of a short helix (approximately residues 976–982) that is completely shielded by the RBD when spike is in the closed form, though becomes slightly more exposed in the open conformation. Residue 681 is part of the loop (residues 676–689) containing the furin-cleavage site, the structure of which is disordered in both cleaved and uncleaved forms[27], though the surface-exposed locations of modelled residues 676 and 689 (orange in Fig. 4B) indicate that the unmodelled residues 677–688 form a prominently-exposed loop; the significant structural flexibility of which has prevented inclusion in structural models[22,27] .

Discussion

We have presented data demonstrating multiple, independent, and circulating lineages of SARS-CoV-2 variants bearing a Spike **Δ**H69/V70. This deletion spanning six nucleotides, is mostly due to an out of frame deletion of six nucleotides, has frequently followed receptor binding amino acid replacements (N501Y, N439K and Y453F that have been shown to increase binding affinity to hACE2 and reduce binding with monoclonal antibodies) and its prevalence is rising internationally. Interestingly the presence of sequence at site 69/70 appears to be unique to SARS-CoV-2 and the closest bat sarbecovirus, RaTG13, and one of the pangolin sequences. We speculate it may have been lost in the other pangolins as these viruses are presumably originated in bats infecting the pangolins after importation to China.

The **Δ**H69/V70 deletion was also shown to increase Spike mediated infectivity by two-fold over a single round of infection, and appeared to occur with a mutation that conferred reduced susceptibility to neutralising antibodies[28]. Over the millions of replication rounds per day in a SARS-CoV-2 infection even modest reductions in antibody susceptibility could be significant. Therefore, **Δ**H69/V70 may be a 'permissive' mutation that enhances replication[28], with the potential to enhance the ability of SARS-CoV-2 to generate immune/ vaccine escape variants that would have otherwise significantly reduced viral fitness. In the case of the UK B1.1.7 lineage with multiple Spike mutations that included the key mutations N501Y and **Δ**H69/V70, we were able to detect a sequence basal in the phylogeny of B1.1.7 that had both mutations as well as D1118H in S2.

The potential for SARS-CoV-2 evolve to rapidly emerge and fix mutations is exemplified by D614G, an amino acid replacement in S2 that alters linkages between S1 and S2 subunits on adjacent protomers as well as RBD orientation, infectivity, and transmission[26,29,30]. The example of D614G also demonstrates that mechanisms directly impacting important biological processes can be indirect. Similarly, a number of possible mechanistic explanations may underlie **Δ**H69/V70. For example, the fact that it sits on an

exposed surface and is estimated to alter the conformation of a particularly exposed loop might be suggestive of immune interactions and escape, although allosteric interactions could alternatively lead to the higher infectivity recently reported[31].

The finding of a lineage (B.1.1.7), termed VOC 202012/01, 8bearing seven S gene mutations across the RBD (N501Y, A570D), S1 (ΔH69/V70 and Δ144) and S2 (P681H, T716I, S982A and D1118H) in UK requires urgent experimental characterisation. The detection of a high number of novel mutations suggests this lineage has either been introduced from a geographic region with very poor sampling or viral evolution may have occurred in a single individual in the context of a chronic infection[12]. This variant bears some concerning features: firstly, the ΔH69/V70 deletion which increases infectivity by two fold[31]. Secondly the Δ144 which may affect binding by antibodies related to 4A8[13]. Thirdly VOC 202012/01 bears the N501Y mutation that may have higher binding affinity for ACE2 and which has arisen independently in other countries, including South Africa where it has led to establishment and explosive transmission of a multi-mutated lineage[18]. Finally, the VOC 2020/1201 lineage has a second RBD mutation A570D that could alter Spike RBD structure and a mutation near to the furin cleavage site could represent further adaptative change. The emergence of multi-mutated variants in the UK and South Africa may herald an era of re-infection and threaten future vaccine efficacy if left unchecked.

Given the emergence of multiple clusters of variants carrying RBD mutations and the ΔH69/V70 deletion, limitation of transmission takes on a renewed urgency. Continued emphasis on testing/tracing, social distancing and mask wearing are essential, with investment in other novel methods to limit transmission[32]. In concert, comprehensive vaccination efforts in the UK and globally should be accelerated in order to further limit transmission and acquisition of further mutations. If geographically limited then focussed vaccination may be warranted. Research is vitally needed into whether lateral flow devices for antigen and antibody detection can detect emerging strains and the immune responses to them. Finally, detection of the deletion and other key mutations by rapid diagnostics should be a research priority as such tests could be used as a proxy for antibody escape mutations to inform surveillance at global scale.

Methods

Phylogenetic Analysis

All available full-genome SARS-CoV-2 sequences were downloaded from the GISAID database (http://gisaid.org/)[33] on 26th November. Duplicate and low-quality sequences (>5% N regions) were removed, leaving a dataset of 194,265 sequences with a length of >29,000bp. All sequences were realigned to the SARS-CoV-2 reference strain MN908947.3, using MAFFT v7.475 with automatic flavour selection and the –keeplength –addfragments options[34]. Major SARS-CoV-2 clade memberships were assigned to all sequences using the Nextclade server v0.10 (https://clades.nextstrain.org/).

Maximum likelihood phylogenetic trees were produced using the above curated dataset using IQ-TREE v2.1.2[35]. Evolutionary model selection for trees were inferred using ModelFinder[36] and trees were estimated using the GTR+F+I model with 1000 ultrafast bootstrap replicates[37]. All trees were visualised with Figtree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) and ggtree v1.14.6 rooted on the SARS-CoV-2 reference sequence and nodes arranged in descending order. Nodes with bootstraps values of <50 were collapsed using an in-house script.

To reconstruct a phylogeny for the 69/70 Spike region of the 17 *Sarbecoviruses* examined in Fig 2, Rdp5[38] was used on the codon Spike alignment to determine the region between amino acids 1 and 256 as putatively non-recombinant. A tree was reconstructed using the protein alignment of this region with FastTree (default parameters)[39]. Alignment visualisation was done using BioEdit[40].

Structural modelling

The structure of the post-deletion NTD (residues 14-306) was modelled using I-TASSER[41], a method involving detection of templates from the protein data bank, fragment structure assembly using replica-exchange Monte Carlo simulation and atomic-level refinement of structure using a fragment-guided molecular dynamics simulation. The structural model generated was aligned with the spike structure possessing the pre-deletion conformation of the 69-77 loop (PDB 7C2L[22]) using PyMOL (Schrödinger). Figures prepared with PyMOL using PDBs 7C2L, 6M0J[42], 6ZGE28 and 6ZGG[27].

Declarations

Conflicts of interest

RKG has received consulting fees from UMOVIS lab, Gilead Sciences and ViiV Healthcare, and a research grant from InvisiSmart Technologies.

Acknowledgements

RKG is supported by a Wellcome Trust Senior Fellowship in Clinical Science (WT108082AIA). SAK is supported by the Bill and Melinda Gates Foundation via PANGEA grant: OPP1175094. DLR is funded by the MRC (MC UU 1201412). WH is funded by the MRC (MR/R024758/1). We thank Dr James Voss for the kind gift of HeLa cells stably expressing ACE2. SL is funded by Medical Research Council MC_UU_12014/12

References

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature* **579**, 270–273 (2020).
2. Sungnak, W. *et al.* SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nature medicine* **26**, 681–687 (2020).
3. Thomson, E. C. *et al.* The circulating SARS-CoV-2 spike variant N439K maintains fitness while evading antibody-mediated immunity. *bioRxiv* (2020).
4. Starr, T. N. *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310. e1220 (2020).
5. Greaney, A. J. *et al.* Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host & Microbe* (2020).
6. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *bioRxiv* (2020).
7. Choi, B. *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *New England Journal of Medicine* **383**, 2291–2293 (2020).
8. MacLean, O. A. *et al.* Natural selection in the evolution of SARS-CoV-2 in bats, not humans, created a highly capable human pathogen. *BioRxiv* (2020).
9. consortiumcontact@cogconsortium.uk., C.-G. U. C.-U. An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99-e100, doi:10.1016/s2666-5247(20)30054-9 (2020).
10. Mlcochova, P. *et al.* Combined Point-of-Care Nucleic Acid and Antibody Testing for SARS-CoV-2 following Emergence of D614G Spike Variant. *Cell Reports Medicine* **1**, 100099 (2020).
11. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine* **26**, 672–675, doi:10.1038/s41591-020-0869-5 (2020).
12. Kemp, S. *et al.* Neutralising antibodies drive Spike mediated SARS-CoV-2 evasion. *medRxiv*, 2020.2012.2005.20241927, doi:10.1101/2020.12.05.20241927 (2020).
13. McCarthy, K. R. *et al.* Natural deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *bioRxiv*, 2020.2011.2019.389916, doi:10.1101/2020.11.19.389916 (2020).
14. Lam, T. T.-Y. *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282–285, doi:10.1038/s41586-020-2169-0 (2020).
15. Munnink, B. B. O. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* (2020).
16. Baum, A. *et al.* Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* **369**, 1014–1018, doi:10.1126/science.abd0831 (2020).
17. Gu, H. *et al.* Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* **369**, 1603–1607, doi:10.1126/science.abc4730 (2020).

18. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*, 2020.2012.2021.20248640, doi:10.1101/2020.12.21.20248640 (2020).
19. Weisblum, Y. *et al.* Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife* **9**, e61312, doi:10.7554/eLife.61312 (2020).
20. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295–1310.e1220, doi:10.1016/j.cell.2020.08.012 (2020).
21. Rambaut A., L. N., Pybus O, Barclay W, Carabelli A. C., Connor T., Peacock T., Robertson D. L., Volz E., on behalf of COVID-19 Genomics Consortium UK (CoG-UK). *Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations*, <https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563> (2020).
22. Chi, X. *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).
23. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature medicine* **26**, 450–452 (2020).
24. Davidson, A. D. *et al.* Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome medicine* **12**, 1–15 (2020).
25. Peacock, T. P. *et al.* The furin cleavage site of SARS-CoV-2 spike protein is a key determinant for transmission due to enhanced replication in airway cells. *bioRxiv* (2020).
26. Yurkovetskiy, L. *et al.* Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **183**, 739–751. e738 (2020).
27. Wrobel, A. G. *et al.* SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nature Structural & Molecular Biology* **27**, 763–767, doi:10.1038/s41594-020-0468-7 (2020).
28. Kemp, S. *et al.* Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/V70. *bioRxiv*, 2020.2012.2014.422555, doi:10.1101/2020.12.14.422555 (2020).
29. Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827. e819 (2020).
30. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* (2020).
31. Kemp, S. *et al.* Neutralising antibodies drive Spike mediated SARS-CoV-2 evasion. *medRxiv*, 2020.2012.2005.20241927, doi:10.1101/2020.12.05.20241927 (2020).
32. Mlcochova, P. *et al.* Extended in vitro inactivation of SARS-CoV-2 by titanium dioxide surface coating. *bioRxiv* (2020).

33. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **22**, 30494, doi:10.2807/1560-7917.ES.2017.22.13.30494 (2017).
34. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780, doi:10.1093/molbev/mst010 (2013).
35. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534, doi:10.1093/molbev/msaa015 (2020).
36. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* **14**, 587–589 (2017).
37. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution* **30**, 1188–1195, doi:10.1093/molbev/mst024 (2013).
38. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution* **1** (2015).
39. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**, 1641–1650, doi:10.1093/molbev/msp077 (2009).
40. Hall, T., Biosciences, I. & Carlsbad, C. BioEdit: an important software for molecular biology. *GERF Bull Biosci* **2**, 60–61 (2011).
41. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols* **5**, 725–738 (2010).
42. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220, doi:10.1038/s41586-020-2180-5 (2020).

Figures

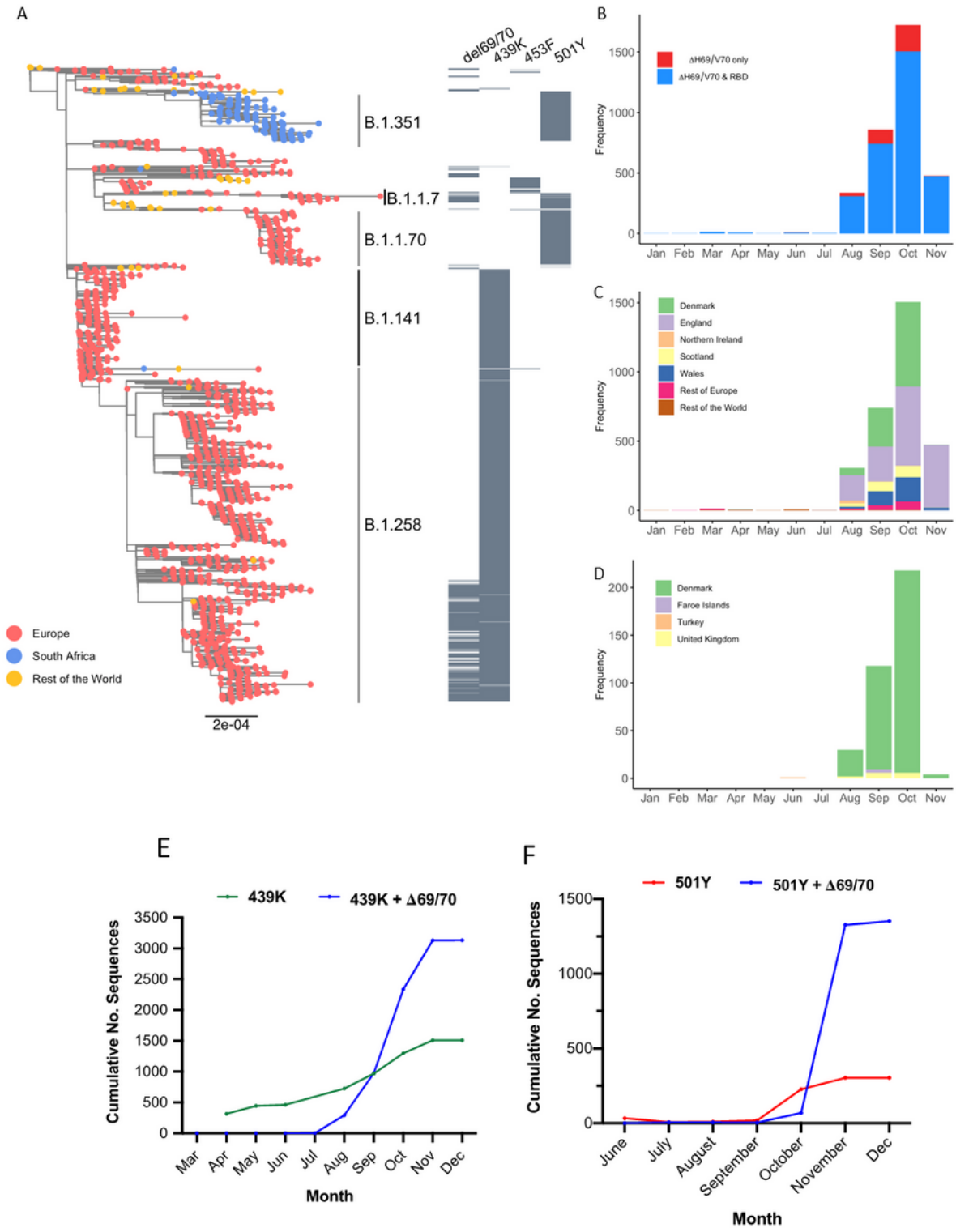

Figure 1

Sub-sampled global phylogeny of SARS-CoV-2 whole genome sequences highlighting those with specific mutations in Spike: ΔH69/V70, N439K, Y453F and N501Y. Tree tips are coloured by geographic region (see key). Grey bars on the right show the presence or absence of the deletion H69/V70 and amino acid variants N439K, Y453F, and N501Y. Lineages from Rambaut et al. 2020 are shown. B–D. Number of new occurrences of SARS-CoV-2 sequences with the ΔH69/V70 deletion. Frequency of the ΔH69/V70

deletion from the GISAID database (accessed 16th Dec 2020) by reporting country and sampling date: B. Worldwide carriage of ΔH69/V70; most sequences preferentially carry RBD mutations alongside the deletion. C. Several distinct ΔH69/V70 lineages carrying RBD mutations 439K, 453F and 501Y have begun to emerge, predominantly in Denmark and England. D. Sequences carrying the ΔH69/V70 but in the absence of other major Spike mutations are restricted mostly to Denmark. The relative increase in frequency of sequences carrying Spike mutants 439K E. 501Y F. and ΔH69/70; viruses co-carrying the Δ69/70 deletion became dominant between September-October 2020 in terms of cumulative cases.

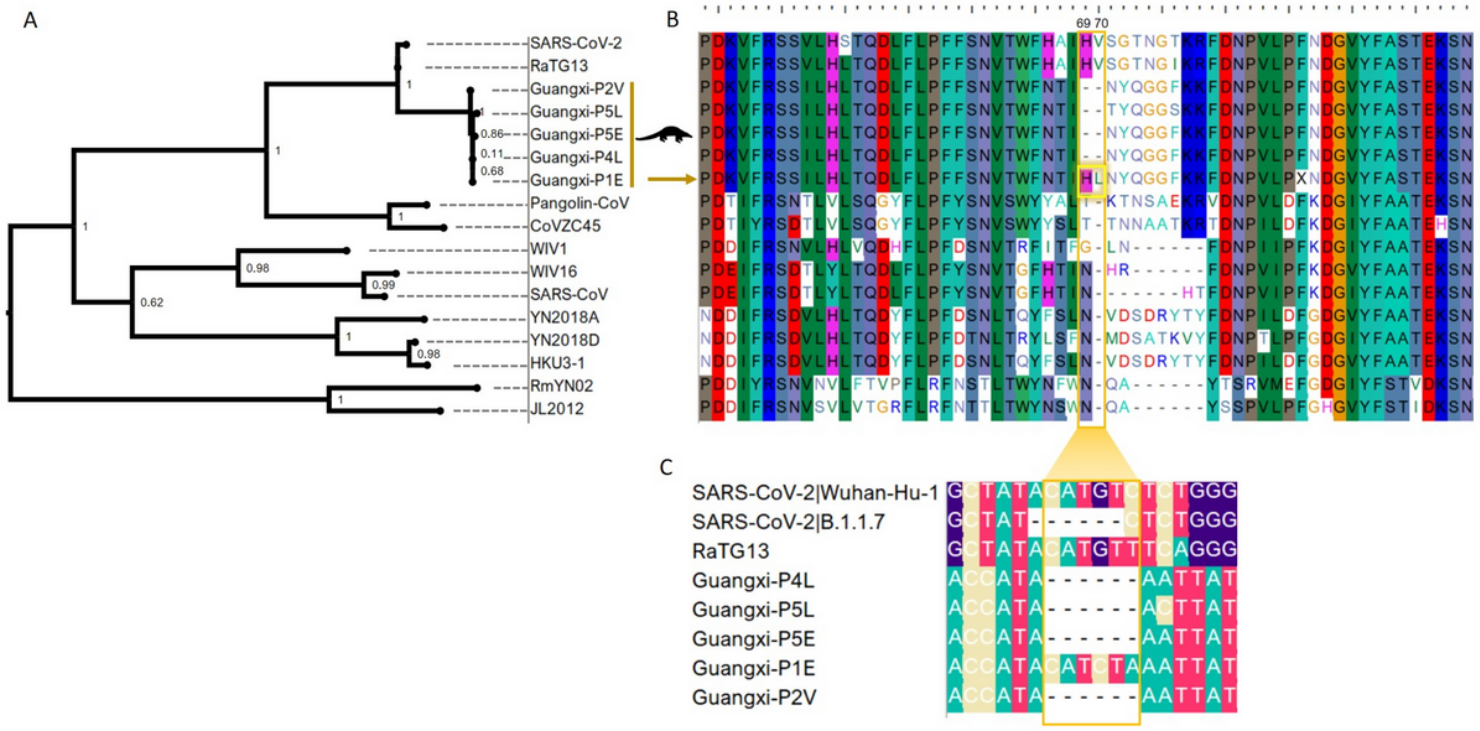

Figure 2

Comparison of the H69/V70 deletion site to other sarbecoviruses. A. phylogeny for the Spike region 1-256 for B. protein sequences from 17 Sarbecoviruses, including SARS-CoV-2 (Wuhan-Hu-1) and SARS-CoV (HSZ-Cc), with distinct genotypes at the Spike region around amino acid positions 69 and 70 (highlighted in orange box). The 69/70 HL insertion in the P1E sequence from the Guangxi pangolin virus cluster is highlighted in a yellow box. C. The nucleotide alignment between SARS-CoV-2 Wuhan-Hu-1, B.1.1.7, the bat sarbecovirus (RatG13) and the Guangxi pangolin viruses shows the difference between the out of frame deletion observed in the former and the in-frame deletion in the latter.

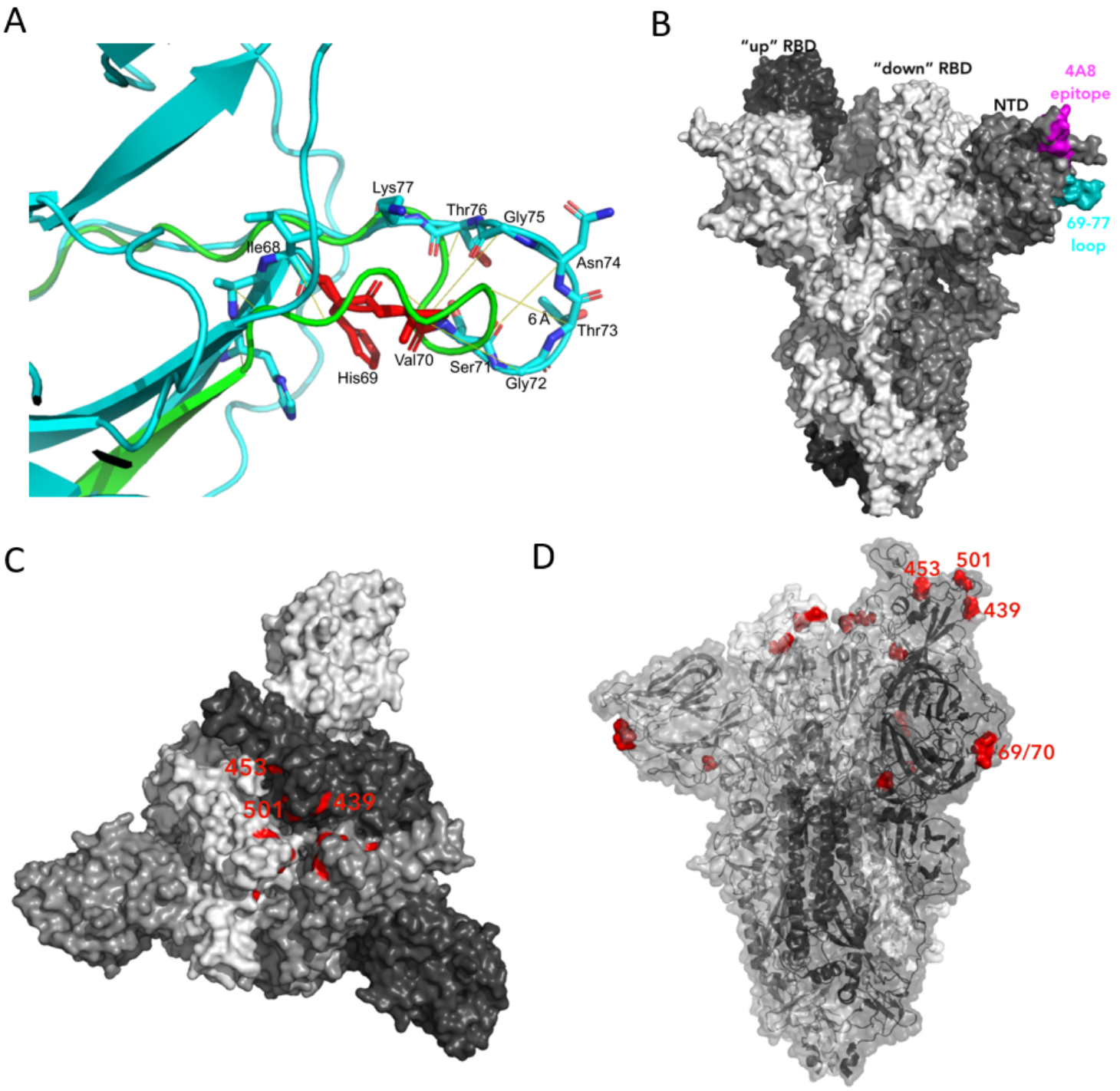

Figure 3

Structural characteristics of viruses carrying deletion at H69/V70 and co-occurring RBD mutations
A. Prediction of conformational change in the spike N-terminal domain due to deletion of residues His69 and Val70. The pre-deletion structure is shown in cyan, except for residues 69 and 70, which are shown in red. The predicted post-deletion structure is shown in green. Residues 66-77 of the pre-deletion structure are shown in stick representation and coloured by atom (nitrogen in blue, oxygen in coral). Yellow lines connect aligned residues 66-77 of the pre- and post-deletion structures and the distance of 6 Å between aligned alpha carbons of Thr73 in the pre- and post-deletion conformation is labelled. B. Surface

representation of spike homotrimer in open conformation (PDB: 7C2L) with each monomer shown in different shades of grey. On the monomer shown positioned to the right, the exposed loop consisting of residues 69-77 is shown in cyan and the neutralising antibody (4A8) binding NTD epitope described by Chi et al. is shown in magenta C. Surface representation of spike in closed conformation (PDB: 6ZGE) viewed in a 'top-down' view along the trimer axis. The residues associated with RBD substitutions N439K, Y453F and N501Y are highlighted in red and labelled on a single monomer. D. Spike in open conformation with a single erect RBD (PDB: 6ZGG) in trimer axis vertical view with the location of ΔH69/V70 in the N-terminal domain and RBD mutations highlighted as red spheres and labelled on the monomer with erect RBD. Residues 71-75, which form the exposed loop undergoing conformational change in A, are omitted from this structure.

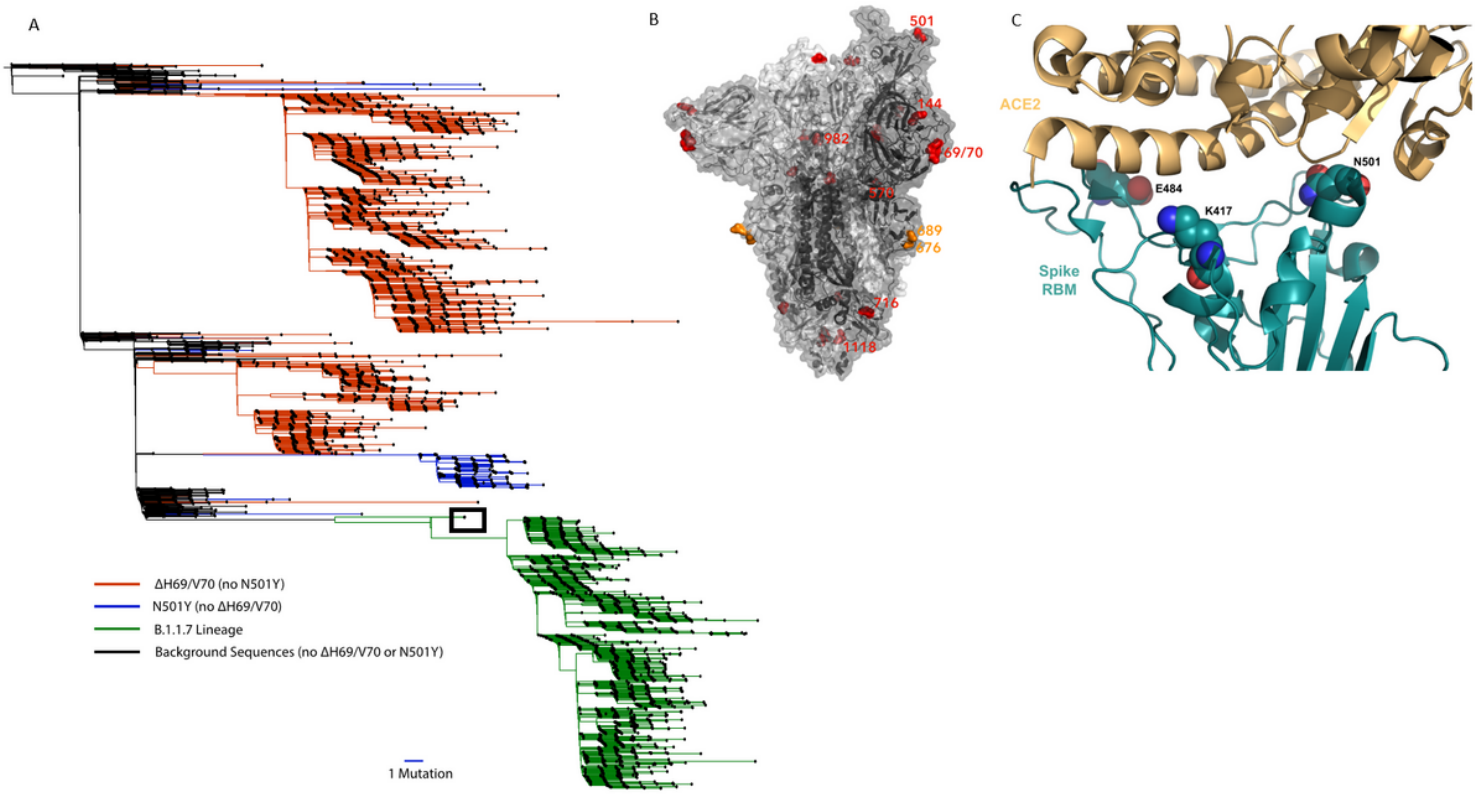

Figure 4

A. Lineages bearing Spike N501Y and ΔH69/V70 mutations: A sub-sampled global phylogeny of SARS-CoV-2 sequences bearing either the ΔH69/V70 or N501Y mutations. Duplicate sequences were removed, and 200 randomly sub-sampled background sequences (without either mutation) were included to create a representative sub-sample of sequences. Distinct sub-lineages of the ΔH69/V70 mutation are currently currently circulating, predominantly in the UK. The novel UK variant VOC 202012/01 (Lineage B.1.1.7, green) is also shown and in black box a related sequence carrying N501Y, ΔH69/V70 and D1118H. B. Surface representation of spike homotrimer in open conformation with one upright RBD overlaid with ribbon representation (PDB: 6ZGG, Wrobel et al., 2020), with different monomers shown in shades of grey. The deleted residues H69 and V70 and the residues involved in amino acid substitutions (501, 570, 716, 982 and 1118) and the deletion at position 144 are coloured red on each monomer and labelled on the

monomer with an upright RBD. The location of an exposed loop containing the furin cleavage site and including residue 681 is absent from the structure, though modelled residues either side of this loop, 676 and 689, are coloured orange. C. Representation of Spike RBM:ACE2 interface (PDB: 6M0J) with residues K417, E484 and N501 highlighted as spheres coloured by element.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- submitted28.12suppfigs.pptx