SARS-CoV-2 variants, spike mutations and immune escape

William T. Harvey[1,2,8], Alessandro M. Carabelli[3,8], Ben Jackson[4], Ravindra K. Gupta[5], Emma C. Thomson[6,7], Ewan M. Harrison[3,7], Catherine Ludden[3], Richard Reeve[1], Andrew Rambaut[4], The COVID-19 Genomics UK (COG-UK) Consortium[*], Sharon J. Peacock[3] and David L. Robertson[2†]

[1]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

[2]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK

[3]Department of Medicine, University of Cambridge, Cambridge, UK

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

[5]Cambridge Institute of Therapeutic Immunology and Infectious Disease (CITIID), University of Cambridge, Cambridge, UK

[6]Department of Clinical Research, London School of Hygiene and Tropical Medicine, London, UK

[7]Wellcome Sanger Institute, Hinxton, UK

[*]A full list of authors and their affiliations appears at the end of the paper.

[8]These authors contributed equally: William T. Harvey, Alessandro M. Carabelli

[†]e-mail: david.l.robertson@glasgow.ac.uk

Abstract

Although most mutations in the SARS-CoV-2 genome are expected to be either deleterious and swiftly purged or relatively neutral, a small proportion will affect functional properties and may alter infectivity, disease severity or interactions with host immunity. Emergence of SARS-CoV-2 in late 2019 was followed by a period of relative evolutionary stasis lasting about eleven months. Since late 2020, however, SARS-CoV-2 evolution has been characterised by the emergence of sets of mutations, in the context of 'variants of concern', that impact virus characteristics including transmissibility and antigenicity, probably in response to the changing immune profile of the human population. There is emerging evidence for reduced neutralisation of some SARS-CoV-2 variants by post-vaccination serum; however, a greater understanding of correlates of protection is required to evaluate how this may impact vaccine effectiveness. Nonetheless, manufacturers are preparing platforms for a possible update to vaccine sequences, and it is crucial that surveillance of genetic and antigenic changes in the global virus population is carried out alongside experiments to elucidate the phenotypic impacts of mutations. In this Review, we summarise the literature on mutations to the SARS-CoV-2 spike protein, the primary antigen, focussing on their impacts on antigenicity and contextualising them in the protein structure, and observed mutation frequencies in global sequence datasets.

[H1] Introduction

As of April 2021, SARS-CoV-2, the causative agent of COVID-19, accounts for over 143 million infections and more than three million deaths worldwide[1]. Virus genomic sequences are being generated and shared at an unprecedented rate, with over one million SARS-CoV-2 sequences available via the Global Initiative on Sharing All Influenza Data (GISAID), permitting near real-time surveillance of the unfolding pandemic[2]. The use of pathogen genomes on this scale to track the spread of the virus internationally, study local outbreaks and inform public health policy signifies a new age in virus genomic investigations[3]. Further to understanding epidemiology, sequencing enables identification of emerging SARS-CoV-2 variants and sets of mutations potentially linked to changes in viral properties.

As highly deleterious mutations are rapidly purged, the majority of mutations observed in genomes sampled from circulating SARS-CoV-2 viruses are expected to be either neutral or mildly deleterious. This is because, although high-effect mutations that contribute to virus adaption and fitness do occur, they tend to be in the minority compared to tolerated low or no-effect 'neutral' amino acid changes[4]. A small minority of mutations are expected to impact virus phenotype in a way that confers a fitness advantage, in at least some contexts. Such mutations may alter various aspects of virus biology such as pathogenicity, infectivity, transmissibility and/or antigenicity. Although care has to be taken not to confound mutations being merely correlated with growing lineages[5], fitness-enhancing mutations were first detected to have arisen within a few months of the evolution of SARS-CoV-2 within the human population. For example, the spike protein amino acid change D614G was noted to be increasing in frequency in April 2020 and to have emerged several times in the global SARS-CoV-2 population and the coding sequence exhibits a high dN/dS ratio **[G]**, suggesting positive selection at the codon[6,7]. Subsequent studies indicated that D614G confers a moderate advantage for infectivity[8,9] and transmissibility[10]. Several other spike mutations of note have now arisen and are discussed in this Review, with a particular focus on mutations affecting antigenicity.

The extent to which mutations affecting the antigenic phenotype of SARS-CoV-2 that will enable variants to circumvent immunity conferred by natural infection or vaccination remains to be determined. However, there is growing evidence that mutations that change the antigenic phenotype of SARS-CoV-2 are circulating and affect immune recognition to a degree that requires immediate attention. The spike protein mediates attachment of the virus to host cell-surface receptors and fusion between virus and cell membranes (**Box 1**)[11]. It is also the principal target of neutralising antibodies generated following infection by SARS-CoV-2 (ref.[12,13]), and is the SARS-CoV-2 component of both mRNA and adenovirus-based vaccines licensed for use and others awaiting regulatory approval[14]. Consequently, mutations that affect the antigenicity of the spike protein are of particular importance. In this Review, we explore the literature on these mutations and their antigenic consequences, focusing on the spike protein and antibody-mediated immunity, and discuss these in the context of observed mutation frequencies in global sequence datasets.

[H1] Spike mutations receiving early attention

The rate of evolution of SARS-CoV-2 from December 2019 to October 2020 was consistent with the virus acquiring approximately two mutations per month in the global population[15,16]. Although our understanding of the functional consequences of spike mutations is rapidly expanding, much of this knowledge involves the reactive investigation of amino acid changes identified as rapidly increasing in frequency or being associated with unusual epidemiological characteristics. Following the emergence of D614G, an amino acid substitution within the receptor-binding motif (RBM), N439K, was noted as increasing in frequency in Scotland in March 2020. Whereas this first lineage with N439K (designated B.1.141 using the Pango nomenclature system[17]) quickly became extinct, another lineage that independently acquired N439K (B.1.258) emerged and circulated widely in many European countries[18]. N439K is noteworthy as it enhances binding affinity to the ACE2 receptor and reduces the neutralising activity of some monoclonal antibodies (mAbs) and polyclonal antibodies present in sera from people who have recovered from infection[18]. Another RBM amino acid change, Y453F — associated with increased ACE2 binding affinity[19] — received considerable attention following its identification in sequences associated with infections in humans and mink; most notably one lineage identified in Denmark and initially named 'cluster 5' (now, B.1.1.298)[20]. As of 5 November 2020, 214 COVID-19 cases of humans infected with SARS-CoV-2 related to mink were all carrying the mutation Y453F (ref. [21]). The B.1.1.298 lineage also has Δ69–70, an amino-terminal domain (NTD) deletion that has emerged several times across the global SARS-CoV-2 population, including in the second N439K lineage, B.1.258. Δ69–70 is predicted to alter the conformation of an exposed NTD loop and has been reported to be associated with increased infectivity[22].

Genomic analyses indicate a change in host environment and signatures of increased selective pressures acting upon immunologically important SARS-CoV-2 genes sampled from around November 2020 (ref.[23]). This coincided with the emergence of variants with higher numbers of mutations, relative to previous circulating variants. These lineages due to their association with increased transmissibility were named 'variants of concern'. They are defined by multiple convergent mutations that are hypothesised to have arisen either in the context of chronic infections or in previously infected individuals[24-29]. In addition to understanding the transmissibility and pathogenicity of these emerging variants, it is crucially important to characterise their antigenicity and the level of cross-protection provided from infection by earlier viruses that are genetically and antigenically similar to the virus that first emerged in December 2019 and which is used in all of the current vaccine preparations. Information on how spike mutations affect antigenic profiles can be derived from structural studies, mutations identified in viruses exposed to mAbs or plasma containing polyclonal antibodies, targeted investigations of variants using site-directed mutagenesis and deep mutational scanning (DMS) experiments that systematically investigate the possibility of mutations arising.

[H1] Immunogenic regions of spike

Several studies have probed the antigenicity of SARS-CoV-2 spike protein by epitope mapping **[G]** approaches including solving the structure of the spike protein in complex with the antigen-binding fragment of particular antibodies[13,30-32]. Serological analyses of almost 650 SARS-CoV-2-infected individuals indicated that ~90% of the plasma or serum neutralising antibody activity targets the spike receptor-binding domain (RBD)[12]. A relative lack of glycan-shielding **[G]** may contribute to the immunodominance **[G]** of the RBD[33]. One study reported structural, biophysical and bioinformatics analyses of 15 SARS-CoV-2 RBD-binding neutralising antibodies[31]. Antibody footprints **[G]** were generated by structural analyses of the spike residues considering potential hydrogen bonds and van der Waals interactions with an mAb atom that were less than 4.0 Å. Structural analyses allowed the categorisation of RBD-binding neutralising antibodies into four classes (**Figure 2a**-**b**): ACE2-blocking antibodies that bind the spike protein in the open conformation (class 1); ACE2-blocking antibodies that bind RBD in both open and closed conformations (class 2); antibodies that do not block ACE2 and bind the RBD both in open and closed conformations (class 3); and neutralising antibodies that bind outside the ACE2 site and only in the open conformation (class 4)[31]. Within the RBD, RBM epitopes **[G]** overlapping the ACE2 site are immunodominant whereas other RBD sites generate lower and variable responses in different individuals[12].

Although the RBD is immunodominant, there is evidence for a substantial role of other spike regions in antigenicity, most notably the NTD[13,30,34]. Early structural characterisation of NTD-specific antibodies 4A8 (ref.[32]) and 4-8 (ref.[13]) revealed similar epitope locations towards the upper side of the most prominently-protruding area of the NTD. Cryo-electron microscopy was used to determine the antibody footprint of the neutralising antibody 4A8 and showed key interactions involving spike residues Y145, H146, K147, K150, W152, R246 and W258 (ref.[32]). Epitope binning **[G]** of 41 NTD-specific mAbs led to the identification of six antigenic sites, one of which is recognised by all known NTD-specific neutralising antibodies and has been termed the 'NTD supersite' consisting of residues 14–20, 140–158 and 245–264 (ref.[30]) (**Figure 2a-b**). The mechanism of neutralisation by which NTD-specific antibodies act remains to be fully determined though may involve the inhibition of conformational changes or proposed interactions with auxiliary receptors such as DC-SIGN or L-SIGN[32,35]. Relatively little is known of antigenicity in the S2 subunit, with immunogenicity thought to be impeded by extensive glycan shielding[36], and although both linear and cross-reactive conformational S2 epitopes have been described[37,38], the biological significance of these is not yet known.

[H1] Spike RBD mutations and immune escape

Several studies have contributed to the current understanding of how mutations in the SARS-CoV-2 spike protein affect neutralisation. These studies include traditional escape mutation work that identify mutations, which emerge in virus populations exposed to either mAbs[39] or

convalescent plasma containing polyclonal antibodies[40,41]; targeted characterisations of particular mutations[18,42]; and wider investigations of either large numbers of circulating variants[43] or of all possible amino acid substitutions in the RBD[39,44-46]. For spike residues where mutations have been shown to influence polyclonal antibody recognition, the observation of an effect on either mAb or plasma is indicated in **Figure 2b**. For a smaller number of residues, escape mutations emerging in virus exposed to mAbs or polyclonal plasma have been described ('mAb emerge' and 'plasma emerge' in **Figure 2b**).

In a DMS study, researchers assessed all possible single amino acid variants using a yeast-display system and detected variants that escape either nine neutralising SARS-CoV-2 mAbs[45] or convalescent plasma from eleven individuals taken at two time points post-infection (shades of green in **Figure 2b**)[39]. The resulting heat maps provide rich data on the antigenic consequence of RBD mutations with the plasma escape mutations being of particular interest given that they impact neutralisation by polyclonal antibodies of the kind SARS-CoV-2 encounters in infections, with significant levels of immunity acquired through prior exposure or vaccination. Although significant inter- and intra-person heterogeneity in the impact of mutations on neutralisation by polyclonal serum has been described, the mutations that reduce antibody binding the most occur at a relatively small number of RBD residues indicating substantial immunodominance within the RBD[39].

Of all RBD residues to which substitutions affected recognition by convalescent sera, DMS identified E484 as being of principal importance with amino acid changes to K, Q or P reducing neutralisation titres by more than an order of magnitude[39]. E484K has also been identified as an escape mutation that emerges during exposure to mAbs C121 and C144 (ref.[40]) and convalescent plasma[41], and was the only mutation described in one study as able to reduce the neutralising ability of a combination of mAbs (REGN10989 and REGN10934) to an unmeasurable level[47]. In an escape mutation study using 19 mAbs, substitutions at E484 emerged more frequently than at any other residue (in response to four mAbs) and each of the four 484 mutants identified (E484A, E484D, E484G and E484K) subsequently conferred resistance to each of four convalescent sera tested[48]. No other mAb-selected escape mutants escaped each of the four sera, though mutations K444E, G446V, L452R and F490S escaped three of the four tested[48].

Mutations at position 477 of the spike protein (S477G, S477N and S477R) rank prominently among mAb escape mutations identified by one study, and the mutation S477G conferred resistance to two of the four sera tested[48]. However, substitutions at 477 were not identified as being important in DMS with convalescent plasma[39]. The mutation N439K increases affinity for ACE2 (ref.[19]), is predicted to result in an additional salt bridge at the RBM–ACE2 interface and is thought to preferentially reduce the neutralisation potential of plasma that already has lower neutralising activity[18]. However, a DMS study[39] did not find that the mutation N439K significantly alters neutralisation by polyclonal antibodies in plasma in contrast to previous studies that found that N439K reduced neutralisation by mAbs and convalescent plasma[18]. One explanation for this inconsistency is that the mechanism of immune escape conferred by

N439K is through increased ACE2 affinity rather than by directly affecting antibody epitope recognition and that perhaps the experimental design of the DMS study is less sensitive to detecting immune evasion mutations of this type.

[H1] Spike NTD mutations and immune escape

In the NTD, the majority of evidence for immune evasion focuses on a region centred on a conformational epitope consisting of residues 140–156 (N3 loop) and 246–260 (N5 loop), which includes the epitope of the antibody 4A8 (**Figure 2**; magenta)[32]. In studies that identified the emergence of antibody escape mutations in virus populations exposed to convalescent plasma, mutations are roughly evenly distributed between the RBD and NTD (**Figure 2b**). One study described the emergence of escape mutations in viruses exposed to convalescent plasma from two individuals, one of which selected for NTD mutations only (N148S, K150R, K150E, K150T, K150Q and S151P)[40]. This was despite the plasma being a source of the highly potent RBD-targeting mAb C144 (ref.[40]). NTD antibody escape mutations were not observed for the other samples of plasma investigated and furthermore, the 148–151 mutants only exhibited marginal reductions in sensitivity to the plasma tested, indicating individual immune responses may be differentially affected by mutations to RBD and NTD epitopes[40].

Deletions in the NTD have been observed repeatedly in the evolution of SARS-CoV-2 and have been described as changing NTD antigenicity[30,41,42]. One study identified four recurrently deleted regions (RDRs) within the NTD and tested five frequently observed deletions within these: Δ69–70 (RDR1), Δ141–144 and Δ146 (RDR2), Δ210 (RDR3) and Δ243–244 (RDR4) [42]. Of the four RDRs, RDR 1, RDR2 and RDR4 correspond to NTD loops N2, N3, N5, whereas RDR3 falls between N4 and N5 in another accessible loop (**Figure 3a**; asterisk). Both RDR2 deletions, Δ141–144 and Δ146, and Δ243–244 (RDR4) abolished binding of 4A8 (ref.[42]). Further evidence of the role of RDR2 deletions in immune escape was provided by a study which describes the emergence of Δ140 in SARS-CoV-2 co-incubated with potently neutralising convalescent plasma, causing a four-fold reduction in neutralisation titre[41]. This Δ140 spike mutant subsequently acquired the E484K mutation resulting in a further 4-fold drop in neutralisation titre, thus a two residue change across NTD and RBD can drastically evade the polyclonal antibody response. The Δ140+E484K double mutant next acquired an 11-residue insertion in the NTD N5 loop between Y248 and L249, completely abolishing neutralisation. This insertion, which also introduced a new glycosylation motif in the vicinity of RDR4, is predicted to alter the structure of the antigenic N3 and N5 NTD loops[41]. This finding further demonstrates the structural plasticity of the NTD and indicates that insertions and the acquisition of additional glycosylation motifs in the NTD are further mechanisms in addition to deletion that lead to the evasion of immunity. Other examples of mutations that impact the epitope–paratope interface indirectly are mutations in the signal peptide region and at cysteine residues 15 and 136, which form a disulfide bond that ‘staples’ the NTD N-terminus against the galectin-like β-sandwich[30]. Mutations at those sites (for example, C136Y and S12P, which alter the

cleavage occurring between residues C15 and V16) have been shown to affect the neutralising activity of several mAbs, likely disrupting the disulfide bond and therefore dislodging the supersite targeted by several antibodies[30].

Across the spike protein, some mutations that confer escape to neutralising mAbs have little impact on serum antibody binding[39,40,44], possibly because those mAbs are rare in polyclonal sera, targeting subdominant epitopes[12,39,44]. Escape mutations emerging in viruses exposed to convalescent plasma have been identified in both the NTD (ΔF140, N148S, K150R, K150E, K150T, K150Q and S151P) and the RBD (K444R, K444N, K444Q, V445E and E484K) (**Figure 2b**)[40,41]. Notably, mutations emerging under selective pressure from convalescent plasma may be different to those selected by the most frequent mAb isolated from the same plasma[40]. Potentially, observed differences arise because mutations selected by convalescent plasma facilitate escape from multiple mAbs. Less data on the antigenic effects of S2 mutations exist, though D769H has been described as conferring a decreased susceptibility to neutralising antibodies[24]. Residue 769 is positioned in a surface-exposed S2 loop and D769H was found to arise, in linkage with Δ69–70, in an immunocompromised individual treated with convalescent plasma[24].

[H1] Conformational epitopes in spike

To evaluate potential antigenicity across spike, we analysed the protein using BEpro, a program for the prediction of conformational epitopes based on tertiary structure[49]. This approach calculates a structure-based epitope score, which approximates antibody accessibility for each amino acid position. For each residue, the calculated score accounts for the local protein structure: half sphere exposure measures and propensity scores each depend on all atoms within 8–16 Å of the target residue, with weighting towards closer atoms. Due to this aggregation, calculated scores are relatively insensitive to the effects of single amino acid substitutions. Scores were calculated for the spike protein in both closed and open conformations (**Figure 3**). It has been estimated that ~34% of spike proteins are closed and 27% are open (with the remainder in an intermediate form) following furin cleavage[50]. Scores rescaled between zero and one are plotted for the closed conformation in **Figure 3a** and represented on the structure in **Figure 3b**. A limitation of this approach is that it does not account for glycan shielding of residues and likely overestimates scores at the base of the ectodomain for residues closest to the carboxyl-terminus.

Comparisons with reporting of antibody footprints and the impact of mutations on antigenicity indicates that residues with mutations described as affecting recognition by mAbs or antibodies in convalescent plasma (**Figure 2b**) tend to occur at residues with higher structure-based antibody accessibility scores, compared with other residues belonging to epitope footprints and residues not implicated in antigenicity (**Supplementary Figure 1b**)**.**

Notably, scores for residues with mutations described as affecting plasma antibody recognition are also a little higher on average compared with those described as affecting plasma compared with mAbs only. Epitope scores are particularly high for residues with mutations described as emerging during exposure to convalescent plasma (**Supplementary Figure 1b**)[40,41]. Experimental data on emergence of mutants under selective pressure from polyclonal antibodies is relatively rare, though these trends indicate approaches of this kind can contribute to the ranking of residues at which substitutions are likely to impact the polyclonal antibody response.

Within the RBD, the two areas with high structure-based antibody accessibility scores for the closed spike structure (**Figure 3a;** peaks with consecutive residues with scores >0.8) are centred on residues 444–447 and residues 498–500. These areas are represented as yellow patches near the centre of the top-down view of the spike structure in **Figure 3b**. **Figure 3c** shows that, in general, residues become more accessible and likely to form epitopes when the spike protein is in the open conformation and this is especially true for the RBD, particularly for the upright RBD (**Figure 3c**; yellow). In the open form, residues close to the ACE2 binding site (405, 415, 416, 417, and 468) become much more exposed on both the upright RBD and the clockwise adjacent closed RBD (**Figure 3c**; green). The effect of mutations at these positions is likely to be greater for antibodies belonging to RBD class 1. Residues centred on 444–447 and 498–500 maintain high scores on the upright RBD and are joined by residues in areas 413–417 and 458–465. The only RBD residues that become notably less accessible in the open spike structure are residues 476, 477, 478, 586 and 487 of the closed RBD clockwise adjacent to the upright RBD, which become blocked by the upright RBD (**Figure 3c**; green). Several RBD-specific antibodies are only able to bind the open spike protein (RBD classes 1 and 4 (ref.[31])) and interestingly, it has been observed that D614G makes the spike protein more vulnerable to neutralising antibodies by increasing the tendency for the open conformation to occur[51].

Within the NTD, the highest scoring spike residues in the closed form belong to a loop centred on residues 147–150, which each have scores greater than 0.9 (**Figure 3a**; yellow patch to the extreme right of the structure viewed from the side in **Figure 3b**). This loop, known as the N3 loop, is described as forming key interactions with the neutralising antibody 4A8 (ref. [32]). One study described the structure of five previously unmodelled, protruding NTD loops denoting them N1–5. In addition to N3, high-scoring residues (> 0.7) are found at positions 22–26 (N1), 70 (N2), 173-187 (N4), 207-213 (**Figure 3a**; asterisk) and 247–253 (N5). Structural analysis indicates NTD-binding antibodies are likely able to bind epitopes when the spike protein is in either the closed or open conformation (**Figure 3c**). Outside of the NTD and RBD, the highest scoring residues are 676 and 689 (which lie either side of the loop containing the S1–S2 furin cleavage site, which is disordered in both open and closed conformations[50]), 793–794, 808–812, 1,099–1,100, and 1,139–1,146 (**Figure 3a**). When the spike protein is in the open conformation, increased accessibility results in substantially higher potential epitope scores for S2 residues centred on 850–854, which becomes more accessible on all three spike

monomers (**Figure 3c**) and residues 978–984, which become more accessible on the monomer anti-clockwise adjacent to the RBD-up monomer (**Figure 3c**; blue).

[H1] Structural context of spike mutations

In order to assess the impact of spike mutations and their immunological role in the global SARS-CoV-2 population, we combined structural analyses with the observed frequency of mutations in circulating variants (**Figure 4**). Globally, the highest number of amino acid variants, mapped against reference sequence Wuhan-Hu-1 (MN908947), are recorded at amino acid positions 614, 222 and 18 (**Figure 1a**) (among 426,623 high quality sequences retrieved from GISAID on 3 February 2021 and processed using CoV-GLUE). Residues at positions 614 and 222 have relatively low antibody access scores and are positioned ~50 Å from the receptor-binding site residues when the spike protein is in the open conformation (**Figure 4a-b**). As mentioned above, there is evidence that indicates D614G confers a moderate advantage for infectivity[8,9] and increases transmissibility[10]. The spike amino acid substitution with the second highest frequency is A222V, which is present in the 20A.EU1 SARS-CoV-2 cluster (also designated lineage B.1.177). This lineage has spread widely in Europe and is reported to have originated in Spain[52]. There is no evidence for a notable impact of A222V on virus phenotype (that is, infectivity and transmissibility) and so its increase in frequency is generally presumed to have been fortuitous rather than a selective advantage. The substitution L18F has occurred ~21 times in the global population[53] and is associated with escape from multiple NTD-binding mAbs[30].

Among the 5,106 independent substitutions observed in the spike protein (Box 1), 161 are described as affecting recognition by mAbs or polyclonal antibodies in sera, of which 22 are present in more than 100 sequences. On average, variant frequency is higher at amino acid positions where mutations are described as affecting antibody recognition compared with positions with no described substitutions of antigenic importance (**Supplementary Figure 1a**) and high levels of amino acid substitutions are observed at some amino acid positions where mutations are described as affecting recognition by antibodies in convalescent plasma including positions 439 and 484. This indicates that, generally, the amino acid positions at which antibody escape mutations have been detected *in vitro* tolerate mutations, at least to some degree *in vivo*.

Within the RBD, the positions at which amino acid substitutions are present at the highest frequency are located in close proximity to the RBD–ACE2 interface (**Figure 4**). Of the three RBD amino acid substitutions present in several thousand sequences, N439K and N501Y are described above and N501Y is discussed in more detail below in the context of variants of concern. The other, S477N, is estimated to have emerged at least seven times in the global SARS-CoV-2 population and has persisted at a frequency of between 4–7% of sequences globally since mid-June 2020 (ref.[53]). One study described multiple mAbs that selected for the

emergence of S477N and found this mutant to be resistant to neutralisation by the entire panel of RBD-targeting mAbs that were tested. By contrast, when tested with convalescent serum, neutralisation of the S477N mutant was similar to that of wild-type[48]. In common with N439K and N501Y, S477N results in increased affinity for the ACE2 receptor, though to a lesser extent[19,54]. As described in Box 2, substitutions may facilitate immune escape by increasing receptor-binding affinity independently of any effect that they may have on antibody recognition of epitopes, therefore it is possible that such a mechanism contributes to the impact of S477N on neutralisation. Variant frequency is also moderately high at RBD–ACE2 interface amino acid positions 417, 453 and 446. Of these positions, 446 occurs in a location in the spike structure that is predicted to be highly antigenic and substitutions at this site are described as affecting neutralisation by both mAbs and antibodies present in polyclonal serum[39,43,46,48]. Substitutions at amino acid positions 417 and 453 are described below in the context of variants of concern.

[H1] Variants of interest or concern

In addition to single mutations of note, more heavily mutated SARS-CoV-2 lineages have emerged. Arguably the first variant of interest defined by the presence of several spike mutations and referred to as B.1.1.298 (cluster 5), was detected in Denmark spreading among farmed mink and a small number of people[20]. This lineage is characterised by four amino acid differences, ΔH69–V70, Y453F, I692V and M1229I (**Figure 5**). Of these, the Y453F substitution occurs at a residue within the ACE2 footprint and has been shown using DMS to increase ACE2 affinity[19]. In addition, Y453F has been described as reducing neutralisation by mAbs[47]. In late 2020 and early 2021, the emergence and sustained transmission of lineages with mutations that affect the characteristics of the virus received much attention, most notably the lineages B.1.1.7, B.1.351 and P.1 (also known as 501Y.V1, 501Y.V2 and 501Y.V3, respectively). The locations of the spike mutations in B.1.1.298 and in the three lineages (B.1.1.7, B.1.351 and P.1) are annotated (**Figure 5a**) and information on the structural context and consequences of mutations for antibody recognition and ACE2 binding are shown (**Figure 5b**).

Of the lineages summarised in **Figure 5**, several amino acid substitutions are convergent having arisen independently in different lineages: N501Y which is present in lineages B.1.1.7, B.1.351 and P.1; E484K which is present in lineages B.1.351 and P.1 and has been detected as emerging within the B.1.1.7 lineage[55]; and ΔH69–V70 in B.1.1.298 and B.1.1.7. Additionally, B.1.351 and P.1 possess alternative amino acid substitutions K417N and K417T, respectively. Further lineages with these mutations have also been identified. For example, an independent emergence of N501Y in the B.1.1.70 lineage, which is largely circulating in Wales. Residue 501 is at the RBD–ACE2 interface (**Figure 3c**) and N501Y has been shown experimentally to result in one of the highest increases in ACE2 affinity conferred by a single RBD mutation[19]. E484 has been identified as an immunodominant spike protein residue, with various substitutions, including E484K, facilitating escape from several mAbs[40,47,48] as well as

antibodies in convalescent plasma[39-41,48]. E484K is estimated to have emerged repeatedly in the global SARS-CoV-2 population[53] and has been described in two other lineages originating from B.1.1.28 in addition to lineage P.1 reported to be spreading in the Rio de Janeiro state, Brazil (lineage P.2)[56] and in the Philippines (lineage P.3)[57]. Whereas K417 is described in the epitopes of RBD class 1 and 2 antibodies[31], substitutions to K417 tend to affect class 1 antibody binding and are therefore somewhat less important for the polyclonal antibody response to the RBD, which is dominated by class 2 antibody responses that are more susceptible to substitutions such as E484K (refs.[44,58,59]). In addition to their antigenic effect, both K417N and K417T are expected to moderately decrease ACE2 binding affinity[19] (**Figure 5b**). The ΔH69–V70 deletion has been identified in variants associated with immune escape in immunocompromised individuals treated with convalescent plasma[24]. Experiments show that ΔH69–V70 does not reduce neutralisation by a panel of convalescent sera; however, it may compensate for infectivity deficits associated with affinity-boosting RBM mutations that may emerge due to immune-mediated selection[22].

The first genomes belonging to the lineage B.1.1.7 were sequenced in the south of England in September 2020. Lineage B.1.1.7 is defined by the presence of 23 nucleotide mutations across the genome that map to a single branch of the phylogenetic tree[3]. Of these 23 mutations, 14 encode amino acid changes and three are deletions, including six amino acid substitutions in the spike protein (N501Y, A570D, P681H, T716I, S982A and D1118H) and two NTD deletions (ΔH69–V70 and ΔY144)[3]. The lineage has been associated with a rapidly increasing proportion of reported SARS-CoV-2 cases and phylogenetic analyses indicate that this lineage was associated with a growth rate estimated to be 40–70% higher than other lineages[60,61]. There is also evidence that this lineage may be associated with a higher viral load[62]. In addition to N501Y, for which there is some evidence that it reduces neutralisation by a small proportion of RBD antibodies[63], there is evidence for an antigenic effect of ΔY144 (**Figure 5b**). This deletion is expected to alter the conformation of the N3 NTD loop (amino acid positions 140–156) and has been demonstrated to abolish neutralisation by a range of neutralising antibodies[30]. The B.1.1.7 spike mutations have been shown to diminish neutralisation of a higher proportion of NTD-specific neutralising antibodies (9 of 10; 90%) than RBD-specific neutralising antibodies (5 of 31; 16%)[63]. Given the immunodominance of RBD, this could explain the modest reductions in neutralising activity of convalescent sera against authentic B.1.1.7 or pseudoviruses carrying the B.1.1.7 spike mutations[64,65]. The co-occurrence of ΔY144 and E484K is concerning with respect to the polyclonal antibody response as the N3 loop that ΔY144 changes is predicted to be among the most immunogenic regions of spike (**Figure 3a**) and amino acid substitutions at position 484 diminish neutralisation by a range of RBD-targeting antibodies.

Reports of lineages with N501Y circulating in the UK were followed by reports of another lineage possessing N501Y circulating in South Africa (lineage B.1.351), which has been rapidly expanding in frequency since December 2020 (ref.[66]). In addition to N501Y, lineage B.1.351 is defined by the presence of five further spike amino acid substitutions (D80A, D215G, K417N, E484K and A701V) and a deletion in the NTD, Δ242–244. High numbers of B.1.351 viruses also

have the spike amino acid substitutions L18F, R246I and D614G. A similar NTD deletion, Δ243–244, abolishes binding by the antibody 4A8 (ref.[42]), and L18F and R246I also occur within the NTD supersite and likely affect antibody binding as well[30]. The co-occurrence of K417N, E484K and these NTD substitutions suggests that B.1.351 may overcome the polyclonal antibody response by reducing neutralisation by class 1 and 2 RBD-specific antibodies and NTD-specific antibodies (**Figure 5b**). Data reported by one study showed that nearly half of examined convalescent plasma samples (21 of 44; 48%) had no detectable neutralisation activity against the B.1.351 variant[58]. Other experiments with pseudotyped viruses showed that the B.1.351 variant was also resistant to the neutralising activity of some mAbs (12 of 17; 70%)[67]. Similarly, a study showed that the neutralising effect of convalescent plasma collected from fourteen individuals was strongly reduced against a live (authentic) B.1.351 virus (with IC_{50} **[G]** reduced by 3.2–41.9-fold relative to first wave virus)[68].

The P.1 lineage, a sub-lineage of B.1.1.28, was first detected in Brazil[69] and in travellers from Brazil to Japan[70]. Lineage P.1 is characterised by the presence of several amino acid substitutions in the spike protein: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y and T1027I (ref.[69]). In addition to substitutions at positions 417, 484 and 501 discussed above, the P.1 lineage has a cluster of substitutions in close proximity to the described antigenic regions of the NTD including L18F, which is known to reduce neutralisation by some antibodies[30]. The substitutions T20N and P26S also occur in or near the NTD supersite[30] at residues with high antibody accessibility scores (**Figure 5b**) and T20N introduces a potential glycosylation site which could result in glycan shielding (**Box 2**) of part of the supersite. The P.1 lineage has also been associated with a reinfection case in Manaus, Brazil[27], indicating it is contributing to antigenic circumvention of what might have been an otherwise effective immune response. More details of frequency and geographic distribution of the P1 lineage can be found at the Pango lineages website[71].

Increasingly, lineages possessing independent occurrences of mutations in common with the variants of concern B.1.1.7, B.1.351 and P.1 are being detected, demonstrating convergent evolution. For example, viruses of lineage B.1.525 which has been observed in several countries, albeit at low frequency to date, have NTD deletions ΔH69–V70 and ΔY144 in common with viruses of the B.1.1.7 lineage; E484K in common with the B.1.351 and P.1 lineages; and spike amino acid substitutions Q52R, Q677H and F888L (ref.[72]). Repeated amino acid substitutions at position 677 and the independent emergence of Q677H in several lineages in the US provides strong evidence of adaptation, potentially through an effect of this mutation on the proximal polybasic furin cleavage site, though further experiments are required to determine its impact[73]. Other novel variants have been identified spreading in California and New York, US (B.1.427/B.1.429 and B.1.526, respectively). The B.1.427/B.1.429 variants carry an antigenically noteworthy substitution L452R[74], which has been shown to reduce neutralisation by several mAbs[43,45,48,59] and convalescent plasma[43]. L452R has independently appeared in several other lineages across the globe between December 2020 and February 2021, indicating that this amino acid substitution is probably the result of viral adaptation due to increasing immunity in the population[74]. L452R is also present in the A.27

lineage associated with a cluster of cases identified on the island of Mayotte[75]. The lineage B.1.526 has been found to carry either S477N or E484K among other lineage defining mutations[76,77], both of which are described as antigenically important above. A new variant carrying E484K currently designated A.VOI.V2 has been recently identified in Angola from cases involving travel from Tanzania[78]. This variant carries several amino acid substitutions in the spike protein and three deletions in the NTD, some of which are within the antigenic supersite[78]. Another variant within A lineage, the prevalence of which is rising in Uganda (A.23.1), shares with the B.1.1.7 lineage a substitution at position 681 within the furin cleavage site (P681R has been found in the A lineage whereas P681H in the B.1.1.7 lineage) and additionally has the amino acid substitutions R102I, F157L, V367F and Q613H. Q613H is speculated to be important as it occurs at a position neighbouring D614G (ref.[79]). Amino acid position 157 has been identified as an epitope residue with F157A reducing neutralisation by the mAb 2489 (ref.[34]).

New variants will continue to emerge and although it is important to understand the phenotypes of emerging variants in terms of infectivity, transmissibility, virulence and antigenicity, it is also important to quantify the phenotypic impacts of specific mutations present in variants, both individually and in combination with other mutations. As new variants with unforeseen combinations of mutations continue to emerge, such insights will allow predictions of virus phenotype. For example, recently detected viruses of lineage B.1.617.1 were anticipated to show altered antigenicity due to the presence of substitutions L452R and E484Q, which have been described as affecting antibody recognition [39,43,45,48,80]. Moving forward, the experimental characterisation of SARS-CoV-2 spike mutations generated to date will continue to provide extremely useful information on individual mutations or combinations of mutations that may not yet have been seen in circulating viruses.

[H1] Vaccine efficacy against new variants

To date, several vaccines have been licensed and rolled out very successfully in several countries, but the number of individuals vaccinated still represents a small fraction of the global population (**Supplementary Table 1**). In order to assess the impacts of mutations on vaccine efficacy, authentic and pseudoviruses possessing particular spike mutations (either individually or in combination), and larger sets of mutations representing variants of concern and other circulating spikes have been assessed using neutralisation assays with post-vaccination sera (**Supplementary Table 1**). Typically, studies report a fold change in variant virus, or pseudovirus, neutralisation relative to wild-type virus (the serum concentration at which 50% neutralisation (IC_{50}) is achieved with the variant divided by the average IC_{50} for the wild-type virus).

Post vaccination sera from a cohort of 20 volunteers immunised with the mRNA vaccines mRNA-1273 (Moderna) or BNT162b2 (Pfizer–BioNTech) showed high binding titres for anti-SARS-CoV-2 spike IgM and IgG with plasma neutralising activity and relative numbers of RBD-

specific antibodies equivalent to natural infection[59]. Furthermore, epitope mapping of mAbs isolated from post-vaccination sera showed they targeted a range of RBD epitopes similar to those isolated from naturally infected individuals[59]. The plasma neutralising activity and the numbers of RBD-specific memory B cells were found to be equivalent to plasma from individuals who recovered from natural SARS-CoV-2 infection[59]. Investigations with pseudoviruses possessing RBD mutations carried by variants of concern demonstrated that the neutralising activity of plasma from vaccinated individuals showed a small but significant decreases of 1–3-fold against E484K, N501Y or the K417N + E484K + N501Y triple mutant[59]. Other data indicate that the effect of N501Y alone on neutralisation is relatively modest, and other studies using sera from 20 participants in a trial of the BNT162b2 vaccine showed equivalent neutralising titres to pseudoviruses carrying the N501 and Y501 mutations[81]. Other investigations with recombinant viruses carrying either N501Y, ΔH69–V70 + N501Y + D614G or E484K + N501Y + D614G demonstrated that, compared with the Wuhan-Hu-1 reference virus, only E484K + N501Y + D614G resulted in a small and modest reduction in neutralisation by post-vaccination sera elicited by two BNT162b2 doses, and only modest differences in neutralisation were seen compared with the Wuhan-Hu-1 reference virus[82].

As stated above, convalescent plasma from individuals infected with pre-B.1.1.7 viruses (that is, viruses that circulated prior to the emergence of the B.1.1.7 lineage) shows only a modest reduction in neutralisation activity against B.1.1.7 or pseudovirus possessing B.1.1.7 spike mutations[63,77], and results with post-vaccination sera are broadly consistent with this. Pseudoviruses carrying the set of B.1.1.7 spike mutations evaluated with post-vaccination serum from individuals who received the BNT162b2 vaccine (2 doses)[63,77,83] or mRNA-1273 vaccine (2 doses)[63] exhibited only a modest reduction in neutralisation titres (less than 3-fold). However, assays using pseudovirus carrying B.1.1.7 spike mutations and the with the addition of E484K, a combination that has been observed in sequencing of circulating isolates, showed larger, more significant drops (6.7-fold) in neutralisation with post-vaccination sera isolated from individuals who received the BNT162b2 vaccine[84]. In a live-virus neutralisation assay, neutralising titres of ChAdOx1 nCoV-19 (Oxford–AstraZeneca) post-vaccination sera were nine times lower than titres against the B.1.1.7 relative to a canonical non-B.1.1.7 lineage (Wuhan-Hu-1 with the S247R spike mutation)[85]. Similarly, neutralising activity of sera elicited by the inactivated vaccine Covaxin (Bharat Biotech) against B.1.1.7 viruses was largely preserved[86]. Pseudovirus and live-virus neutralisation assays showed that the neutralising activity of sera from individuals after the two doses of the ChaAdOx1 vaccine against the B.1.351 variant was reduced or abrogated[85]. Post-vaccination sera from individuals who received two doses of mRNA-1273 (28 days apart) showed reduced neutralisation against the B.1.351 variant (6.4-fold reduction)[87]. By contrast, neutralising activity of sera elicited by inactivated vaccine BBIBP-CorV (Sinopharm) against the authentic virus B.1.351 showed only a slight reduction (less than 2-fold)[88].

Comparison of the differing extents to which variants affect neutralisation by post-vaccination serum is complicated by the different methodologies used by various studies. However, one

study tested eight SARS-CoV-2 variants of interest or concern including B.1.1.298, B.1.1.7 and P.1 as well as three B.1.351 variants, distinguished by their combination of NTD mutations, representing sequence diversity in circulating viruses of this lineage. Pseudoviruses were generated by the same system and tested with post-vaccination sera from individuals who received two doses of either the BNT-162b2 (n = 30) or mRNA-1273 (n = 35) vaccines[89]. Compared with wild-type, pseudoviruses with D614G, or the mutations defining lineages B.1.1.7, B.1.1.298 and B.1.429 each showed non-statistically significant decreases in neutralisation[89]. Lineages P.1 and P.2 each showed significant decreases with both BNT162b2 (6.7 and 5.8-fold, respectively) and mRNA-1273 (4.5 and 2.9-fold, respectively) post-vaccination sera[89]. The three B.1.351 variants investigated, representing the majority of deposited B.1.351 sequences, showed much larger decreases in neutralisation activity ranging from 34–42-fold (BNT162b2) and 19.2–27.7-fold (mRNA-1273). Taken together, these data indicate that E484K is the primary determinant of the decreases in neutralisation titres, which distinguish P.1, P.2 and the three B.1.351 variants from the other pseudoviruses tested. In addition to E484K, further mutations that are shared by each of the three B.1.351 variants, but are not possessed by P.1. and P.2 lineages, are D80A, Δ242–244, K417N (though K417T is present in P.1) and A701V.

To complement the experimental data provided by neutralisation assays, there is emerging evidence from clinical trials on the impact of variants on vaccine efficacy. Early indications are that these are broadly consistent with the laboratory results, with the B.1.351 variant showing greater signs of vaccine escape. The ChAdOx1 vaccine showed clinical efficacy against the B.1.1.7 variant but failed to provide protection against mild-to-moderate disease caused by the B.1.351 variant, with vaccine efficacy against the variant estimated at 10.4% (95% confidence interval; -76.8 to 54.8)[84,85,90]. Preliminary data from clinical trials reported that the NVX-CoV2373 (Novavax) protein-based vaccine provides 95.6% efficacy against the wild-type virus and that this is moderately lower for B.1.1.7 (85.6%) and further reduced for B.1.351 (60.0%) variants[90]. Similarly, the single dose vaccine JNJ-78436735 (Johnson & Johnson–Janssen) has been shown to provide 72% protection against moderate to severe SARS-CoV-2 infections in the US, but the proportion significantly decreased to 57% in South Africa (at a time when B.1.351 was widespread)[91]. These data indicate that NVX-CoV2373 and JNJ-78436735 are clinically efficacious against the B.1.1.7 and variants circulating in the US but are consistent in that B.1.351 is associated with a larger reduction in vaccine efficacy.

In addition to vaccine efficacy evaluation against SARS-CoV-2 variants and mutations, the effect of mutations on some mAbs used as therapeutics have been described (**Supplementary Table 2**). Single mAb treatment can exert a selective pressure that potentially increases the possibility of mutational escape of the targeted antigen. The risk is likely to be reduced with the use of cocktails of two or more mAbs targeting non-overlapping epitopes. REGN-COV2 (Regeneron) (included in the RECOVERY trial in the UK) and AZD7742 (AstraZeneca) are two examples of mAbs cocktails that have been developed[92]. Importantly, some mutations in the

RBM have already been identified in variants, which are circulating in the UK (for example, N439K, T478I and V483I) and are likely to impact antigenicity.

[H1] Conclusions

There is now clear evidence of the changing antigenicity of the SARS-CoV-2 spike protein and of the amino acid changes that affect antibody neutralisation. Spike amino acid substitutions and deletions that impact neutralising antibodies are present at significant frequencies in the global virus population and there is emerging evidence of variants exhibiting resistance to antibody-mediated immunity elicited by vaccines. Greater understanding of the correlates of immune protection is required to provide context for the results of studies reporting changes in neutralisation. A comprehensive understanding of the consequences of spike mutations on antigenicity will encompass both T cell-mediated immunity and non-spike epitopes recognised by antibodies. To monitor vaccine efficacy and to better understand the implications of antigenic variation for vaccine effectiveness it will be important to collect information on vaccine status and viral genome sequence data from individuals infected with SARS-CoV-2. More generally, a broader understanding of the phenotypic impacts of mutations across the SARS-CoV-2 genome and their consequences for variant fitness will help elucidate drivers of transmission and evolutionary success.

Recent studies have shown the potential selective pressure exerted by convalescent plasma and mAbs treatments on SARS-CoV-2 evolution in immunocompromised individuals[24-26]. Such circumstances, involving long-term virus shedders, may have contributed to the sporadic emergence of the more heavily mutated variants (for example, seen in the B.1.1.7 and B.1.351 lineages). Given that therapeutics (vaccines and antibody-based therapies) mainly target the SARS-CoV-2 spike protein, the selection pressures that favour emergence of new variants carrying immune escape mutations generated in chronic infections[24-26] will be similar to those selecting for mutations that allow reinfections within the wider population[27-29]. Therefore, sequencing of viruses associated with prolonged infections will provide useful information on mutations that could contribute to elevated transmissibility or escape from vaccine-mediated immunity.

The collective data on the effect of mutations on vaccines and convalescent sera efficacy shows that the polyclonal antibody response is focused on a few immunodominant regions, indicating the high probability of future mutation-mediated escape from host immunity. As antigenically different variants are continuing to emerge, it is going to become necessary to routinely collect serum samples from vaccinated individuals and from individuals that have been infected with circulating variants of known sequence. Cross-reactive immunity between circulating lineages can be assessed by measuring the ability of sera to neutralise panels of circulating viruses. The systematic surveillance of antigenic SARS-CoV-2 variants will be enhanced by the establishment of a network similar to the WHO-coordinated Global Influenza Surveillance and Response System (GISRS), a collaborative global effort responsible for tracking the antigenic evolution of human influenza viruses and making recommendations on

vaccine composition. Modelling approaches to predict the evolutionary trajectories of emerging variants based on an understanding of the phenotypic effects of mutations will assist this process, as is the case for the influenza virus[93].

Prediction of the mutational pathways by which a virus such as SARS-CoV-2 will evolve is extremely challenging. Nonetheless, there is a rapidly expanding knowledge base regarding the effect of SARS-CoV-2 spike mutations on antigenicity and other aspects of virus biology. The integration of these data and emerging SARS-CoV-2 sequences has the potential to facilitate the automated detection of potential variants of concern at low frequency, that is, before they are spreading widely. Tracking the emergence of these viruses flagged as potential antigenically significant variants will help to guide the implementation of targeted control measures and further laboratory characterisation. An important part of this process will be the preparation of updated vaccines tailored to emerging antigenic variants that are maximally cross-reactive against all circulating variants. All of these processes will benefit from close international collaboration and the rapid and open sharing of data.

References

1 WHO. *Coronavirus (COVID-19) Dashboard*. https://covid19.who.int/ (2021).

2 Meredith, L. W. *et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect. Dis.* **20**, 1263-1272, https://doi.org/10.1016/S1473-3099(20)30562-4 (2020).

3 Rambaut, A. *et al. 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).

4 Frost, S. D. W., Magalis, B. R. & Kosakovsky Pond, S. L. Neutral Theory and Rapidly Evolving Viral Pathogens. *Mol. Biol. Evol.* **35**, 1348-1354, https://doi.org/10.1093/molbev/msy088 (2018).

5 MacLean, O. A., Orton, R. J., Singer, J. B. & Robertson, D. L. No evidence for distinct types in the evolution of SARS-CoV-2. *Virus Evol.* **6**, https://doi.org/10.1093/ve/veaa034 (2020).

6 Zhan, X.-Y. *et al.* Molecular Evolution of SARS-CoV-2 Structural Genes: Evidence of Positive Selection in Spike Glycoprotein. *bioRxiv*, https://doi.org/10.1101/2020.06.25.170688 (2020).

7 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, https://doi.org/10.1016/j.cell.2020.06.043 (2020).

8 Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464, https://doi.org/10.1126/science.abe8499 (2020).

9 Yurkovetskiy, L. *et al.* Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183**, 739-751.e738, https://doi.org/10.1016/j.cell.2020.09.032 (2020).

10 Volz, E. *et al.* Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *Cell*, https://doi.org/10.1016/j.cell.2020.11.020 (2020).

11 Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562-569, https://doi.org/10.1038/s41564-020-0688-y (2020).

12 Piccoli, L. *et al.* Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* **183**, 1024-1042 e1021, https://doi.org/10.1016/j.cell.2020.09.037 (2020).

13 Liu, L. *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450-456, https://doi.org/10.1038/s41586-020-2571-7 (2020).
14 Dai, L. & Gao, G. F. Viral targets for vaccines against COVID-19. *Nat. Rev. Immunol.* **21**, 73-82, https://doi.org/10.1038/s41577-020-00480-0 (2021).
15 Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* **6**, https://doi.org/10.1093/ve/veaa061 (2020).
16 Worobey, M. *et al.* The emergence of SARS-CoV-2 in Europe and the US. *bioRxiv*, https://doi.org/10.1101/2020.05.21.109322 (2020).
17 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403-1407, https://doi.org/10.1038/s41564-020-0770-5 (2020).
18 Thomson, E. C. *et al.* Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184**, 1171-1187 e1120, https://doi.org/10.1016/j.cell.2021.01.037 (2021).
19 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, https://doi.org/10.1016/j.cell.2020.08.012 (2020).
20 Fonager J. *et al.* Working paper on SARS-CoV-2 spike mutations arising in Danish mink, their spread to humans and neutralization data. https://files.ssi.dk/Mink-cluster-5-short-report_AFO2. (2020).
21 ECDC. *Detection of new SARS-CoV-2 variants related to mink*. https://www.ecdc.europa.eu/sites/default/files/documents/RRA-SARS-CoV-2-in-mink-12-nov-2020.pdf (2020).
22 Kemp, S. *et al.* Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/V70. *bioRxiv*, https://doi.org/10.1101/2020.12.14.422555 (2020).
23 Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *medRxiv*, https://doi.org/10.1101/2021.02.23.21252268 (2021).
24 Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature*, https://doi.org/10.1038/s41586-021-03291-y (2021).
25 Choi, B. *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Eng. J. Med.* **383**, 2291-2293, https://doi.org/10.1056/NEJMc2031364 (2020).
26 Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **183**, 1901-1912 e1909, https://doi.org/10.1016/j.cell.2020.10.049 (2020).
27 Naveca, F. *et al. SARS-CoV-2 reinfection by the new Variant of Concern (VOC) P.1 in Amazonas, Brazil*. https://virological.org/t/sars-cov-2-reinfection-by-the-new-variant-of-concern-voc-p-1-in-amazonas-brazil/596 (2021).
28 Nonaka, C. K. V. *et al. Genomic Evidence of a Sars-Cov-2 Reinfection Case With E484K Spike Mutation in Brazil*. https://www.preprints.org/manuscript/202101.0132/v1 (2021).
29 Resende, P. C. *et al. Spike E484K mutation in the first SARS-CoV-2 reinfection case confirmed in Brazil, 2020*. https://virological.org/t/spike-e484k-mutation-in-the-first-sars-cov-2-reinfection-case-confirmed-in-brazil-2020/584 (2021).
30 McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell*, https://doi.org/10.1016/j.cell.2021.03.028 (2021).
31 Barnes, C. O. *et al.* SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682-687, https://doi.org/10.1038/s41586-020-2852-1 (2020).
32 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, https://doi.org/10.1126/science.abc6952 (2020).
33 Watanabe, Y. *et al.* Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat. Commun.* **11**, 2688, https://doi.org/10.1038/s41467-020-16567-0 (2020).

34 Suryadevara, N. *et al.* Neutralizing and protective human monoclonal antibodies recognizing the N-terminal domain of the SARS-CoV-2 spike protein. *Cell* **184**, https://doi.org/10.1016/j.cell.2021.03.029 (2021).

35 Soh, W. T. *et al.* The N-terminal domain of spike glycoprotein mediates SARS-CoV-2 infection by associating with L-SIGN and DC-SIGN. *bioRxiv*, 2020.2011.2005.369264, https://doi.org/10.1101/2020.11.05.369264 (2020).

36 Casalino, L. *et al.* Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **6**, 1722-1734, https://doi.org/10.1021/acscentsci.0c01056 (2020).

37 Shrock, E. *et al.* Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science* **370**, eabd4250, https://doi.org/10.1126/science.abd4250 (2020).

38 Zheng, Z. *et al.* Monoclonal antibodies for the S2 subunit of spike of SARS-CoV-1 cross-react with the newly-emerged SARS-CoV-2. *Eurosurveillance* **25**, 2000291, https://doi.org/10.2807/1560-7917.ES.2020.25.28.2000291 (2020).

39 Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476 e466, https://doi.org/10.1016/j.chom.2021.02.003 (2021).

40 Weisblum, Y. *et al.* Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* **9**, https://doi.org/10.7554/eLife.61312 (2020).

41 Andreano, E. *et al.* SARS-CoV-2 escape in vitro from a highly neutralizing COVID-19 convalescent plasma. *bioRxiv*, https://doi.org/10.1101/2020.12.28.424451 (2020).

42 McCarthy, K. R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* **371**, 1139-1142, https://doi.org/10.1126/science.abf6950 (2021).

43 Li, Q. *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182**, 1284-1294.e1289, https://doi.org/10.1016/j.cell.2020.07.012 (2020).

44 Greaney, A. J. *et al.* Mutational escape from the polyclonal antibody response to SARS-CoV-2 infection is largely shaped by a single class of antibodies. *bioRxiv*, 2021.2003.2017.435863, https://doi.org/10.1101/2021.03.17.435863 (2021).

45 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* **29**, 44-57 e49, https://doi.org/10.1016/j.chom.2020.11.007 (2021).

46 Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371**, 850, https://doi.org/10.1126/science.abf9302 (2021).

47 Baum, A. *et al.* Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science*, eabd0831, https://doi.org/10.1126/science.abd0831 (2020).

48 Liu, Z. *et al.* Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe* **29**, 477-488 e474, https://doi.org/10.1016/j.chom.2021.01.014 (2021).

49 Sweredoski, M. J. & Baldi, P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. *Bioinformatics (Oxford, England)* **24**, 1459-1460, https://doi.org/10.1093/bioinformatics/btn199 (2008).

50 Wrobel, A. G. *et al.* SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat. Struct. Mol. Biol.* **27**, 763-767, https://doi.org/10.1038/s41594-020-0468-7 (2020).

51 Weissman, D. *et al.* D614G Spike Mutation Increases SARS CoV-2 Susceptibility to Neutralization. *Cell Host Microbe* **29**, 23-31.E24, https://doi.org/10.1016/j.chom.2020.11.012 (2021).

52 Hodcroft, E. B. *et al.* Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv*, https://doi.org/10.1101/2020.10.25.20219063 (2020).

53 Nextstrain. *Nextstrain [accessed on Feb 13 2021]*. https://nextstrain.org/ncov/global?c=gt-S_477&gmax=24271&gmin=22482 (2021).
54 Zahradník, J. *et al.* SARS-CoV-2 RBD *in vitro* evolution follows contagious mutation spread, yet generates an able infection inhibitor. *bioRxiv*, https://doi.org/10.1101/2021.01.06.425392 (2021).
55 Wise, J. Covid-19: The E484K mutation and the risks it poses. *Br. Med. J.* **372**, n359, https://doi.org/10.1136/bmj.n359 (2021).
56 Voloch, C. M. *et al.* Genomic Characterization of a Novel SARS-CoV-2 Lineage from Rio de Janeiro, Brazil. *J. Virol.* **95**, e00119-00121, https://doi.org/10.1128/JVI.00119-21 (2021).
57 Tablizo, F. A. *et al.* Genome sequencing and analysis of an emergent SARS-CoV-2 variant characterized by multiple spike protein mutations detected from the Central Visayas Region of the Philippines. *medRxiv*, 2021.2003.2003.21252812, https://doi.org/10.1101/2021.03.03.21252812 (2021).
58 Wibmer, C. K. *et al.* SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.*, https://doi.org/110.1038/s41591-021-01285-x (2021).
59 Wang, Z. *et al.* mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature*, https://doi.org/10.1038/s41586-021-03324-6 (2021).
60 ECDC. *Risk related to spread of new SARSCoV-2 variants of concern in the EU/EEA*. https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-risk-related-to-spread-of-new-SARS-CoV-2-variants-EU-EEA.pdf (2020).
61 Volz, E. *et al.* Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv*, https://doi.org/10.1101/2020.12.30.20249034 (2021).
62 Kidd, M. *et al.* S-variant SARS-CoV-2 is associated with significantly higher viral loads in samples tested by ThermoFisher TaqPath RT-QPCR. *J. Infect. Dis.*, https://doi.org/10.1093/infdis/jiab082 (2021).
63 Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature*, https://doi.org/10.1038/s41586-021-03412-7 (2021).
64 Rees-Spear, C. *et al.* The impact of Spike mutations on SARS-CoV-2 neutralization. *bioRxiv*, https://doi.org/10.1101/2021.01.15.426849 (2021).
65 Hu, J. *et al.* Emerging SARS-CoV-2 variants reduce neutralization sensitivity to convalescent sera and monoclonal antibodies. *Cell Mol. Immunol.* **18**, 1061-1063, https://doi.org/10.1038/s41423-021-00648-1 (2021).
66 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*, https://doi.org/10.1101/2020.12.21.20248640 (2020).
67 Li, Q. *et al.* No higher infectivity but immune escape of SARS-CoV-2 501Y.V2 variants. *Cell*, https://doi.org/10.1016/j.cell.2021.02.042 (2021).
68 Cele, S. *et al.* Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature*, https://doi.org/10.1038/s41586-021-03471-w (2021).
69 Faria, N. R. *et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings*. https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586 (2021).
70 Naveca, F. *et al. Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein*. https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585 (2021).
71 Rambaut, A. *et al. Global Report Investigating Novel Coronavirus Haplotypes*. https://cov-lineages.org/global_report.html (2020).

72 O'Toole, Á. *Proposal for new lineage within B.1 #4: B.1.525, covlineages / pango-designation.* https://github.com/cov-lineages/pango-designation/issues/4 (2021).
73 Hodcroft, E. B. *et al.* Emergence in late 2020 of multiple lineages of SARS-CoV-2 Spike protein variants affecting amino acid position 677. *medRxiv*, https://doi.org/10.1101/2021.02.12.21251658 (2021).
74 Tchesnokova, V. *et al.* Acquisition of the L452R mutation in the ACE2-binding interface of Spike protein triggers recent massive expansion of SARS-Cov-2 variants. *bioRxiv*, https://doi.org/10.1101/2021.02.22.432189 (2021).
75 Wagner, C., Hodcroft, E., Bell, S. M., Neher, R. & Bedford, T. *Resurgence of SARS-CoV-2 19B clade corresponds with possible convergent evolution*. https://virological.org/t/resurgence-of-sars-cov-2-19b-clade-corresponds-with-possible-convergent-evolution/620 (2021).
76 Annavajhala, M. K. *et al.* A Novel SARS-CoV-2 Variant of Concern, B.1.526, Identified in New York. *medRxiv*, https://doi.org/10.1101/2021.02.23.21252259 (2021).
77 West, A. P., Barnes, C. O., Yang, Z. & Bjorkman, P. J. SARS-CoV-2 lineage B.1.526 emerging in the New York region detected by software utility created to query the spike mutational landscape. *bioRxiv*, https://doi.org/10.1101/2021.02.14.431043 (2021).
78 de Oliveira, T. *et al. A novel variant of interest of SARS-CoV-2 with multiple spike mutations detected through travel surveillance in Africa*. https://www.krisp.org.za/publications.php?pubid=330 (2021).
79 Bugembe, D. L. *et al.* A SARS-CoV-2 lineage A variant (A.23.1) with altered spike has emerged and is dominating the current Uganda epidemic. *medRxiv*, https://doi.org/10.1101/2021.02.08.21251393 (2021).
80 Cherian, S. *et al.* Convergent evolution of SARS-CoV-2 spike mutations, L452R, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. *bioRxiv*, https://doi.org/10.1101/2021.04.22.440932 (2021).
81 Xie, X. *et al.* Neutralization of N501Y mutant SARS-CoV-2 by BNT162b2 vaccine-elicited sera. *bioRxiv*, https://doi.org/10.1101/2021.01.07.425740 (2021).
82 Xie, X. *et al.* Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K, and N501Y variants by BNT162b2 vaccine-elicited sera. *Nat. Med.*, https://doi.org/10.1038/s41591-021-01270-4 (2021).
83 Faulkner, N. *et al.* Reduced antibody cross-reactivity following infection with B.1.1.7 than with parental SARS-CoV-2 strains. *bioRxiv*, https://doi.org/10.1101/2021.03.01.433314 (2021).
84 Emary, K. R. W. *et al.* Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 variant of concern 202012/01 (B.1.1.7): an exploratory analysis of a randomised controlled trial. *Lancet*, https://doi.org/10.1016/S0140-6736(21)00628-0 (2021).
85 Madhi, S. A. *et al.* Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.*, https://doi.org/10.1056/NEJMoa2102214 (2021).
86 Sapkal, G. N. *et al.* Neutralization of UK-variant VUI-202012/01 with COVAXIN vaccinated human serum. *bioRxiv*, https://doi.org/10.1101/2021.01.26.426986 (2021).
87 Wu, K. *et al.* mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants. *bioRxiv*, https://doi.org/10.1101/2021.01.25.427948 (2021).
88 Huang, B. *et al.* Neutralization of SARS-CoV-2 VOC 501Y.V2 by human antisera elicited by both inactivated BBIBP-CorV and recombinant dimeric RBD ZF2001 vaccines. *bioRxiv*, https://doi.org/10.1101/2021.02.01.429069 (2021).
89 Garcia-Beltran, W. F. *et al.* Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*, https://doi.org/10.1016/j.cell.2021.03.013 (2021).
90 Mahase, E. Covid-19: Novavax vaccine efficacy is 86% against UK variant and 60% against South African variant. *Br. Med. J.* **372**, n296, https://doi.org/10.1136/bmj.n296 (2021).
91 Mahase, E. Covid-19: Where are we on vaccines and variants? *Br. Med. J.* **372**, n597, https://doi.org/10.1136/bmj.n597 (2021).

92 Matthews, D. B. A cocktail of antibodies for COVID-19 therapy. *Nat. Rev. Immunol.* **20**, 591, https://doi.org/10.1038/s41577-020-00431-9 (2020).
93 Morris, D. H. *et al.* Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology. *Trends Microbiol.* **26**, 102-118, https://doi.org/10.1016/j.tim.2017.09.004 (2018).
94 Singer, J., Gifford, R., Cotten, M. & Robertson, D. L. *CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation*. http://cov-glue.cvr.gla.ac.uk/#/home (2020).
95 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494, https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 (2017).
96 Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215-220, https://doi.org/10.1038/s41586-020-2180-5 (2020).
97 Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S. & Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **369**, 330, https://doi.org/10.1126/science.abb9983 (2020).
98 Kumar, S., Maurya, V. K., Prasad, A. K., Bhatt, M. L. B. & Saxena, S. K. Structural, glycosylation and antigenic variation between 2019 novel coronavirus (2019-nCoV) and SARS coronavirus (SARS-CoV). *VirusDisease* **31**, 13-21, https://doi.org/10.1007/s13337-020-00571-5 (2020).
99 Hoffmann, M., Kleine-Weber, H. & Pohlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol. Cell* **78**, 779-784 e775, https://doi.org/10.1016/j.molcel.2020.04.022 (2020).
100 Benton, D. J. *et al.* Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* **588**, 327-330, https://doi.org/10.1038/s41586-020-2772-0 (2020).
101 Hensley, S. E. *et al.* Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science* **326**, 734-736, https://doi.org/10.1126/science.1178258 (2009).
102 Silver, Z. A. *et al.* Discovery of O-Linked Carbohydrate on HIV-1 Envelope and Its Role in Shielding against One Category of Broadly Neutralizing Antibodies. *Cell Rep.* **30**, 1862-1869.e1864, https://doi.org/10.1016/j.celrep.2020.01.056 (2020).
103 Das, S. R. *et al.* Fitness costs limit influenza A virus hemagglutinin glycosylation as an immune evasion strategy. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1417, https://doi.org/10.1073/pnas.1108754108 (2011).

Acknowledgements

The authors thank all of the researchers who have shared genome data openly via the GISAID Initiative. COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute. W.T.H. is funded by the MRC (MR/R024758/1). R.R. is funded by the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/R012679/1). D.L.R. and E.C.T. are funded by the MRC (MC_UU_12014/12) and acknowledge the support of the G2P-UK National Virology Consortium (MR/W005611/1) funded by the UKRI. DR also acknowledges support of the Wellcome Trust (WT) (220977/Z/20/Z). AR acknowledges the support of the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS).

Author contributions

W.T.H., A.M.C., B.J., R.K.G., E.C.T., E.M.H., C.L., A.R. and D.L.R. researched data for the article. W.T.H., A.M.C., A.R., S.J.P. and D.L.R. contributed substantially to discussion of the content. W.T.H.,

A.M.C. and D.L.R. wrote the article. W.T.H., A.M.C., R.K.G., E.C.T., R.R., S.J.P. and D.L.R. reviewed and/or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Microbiology thanks Y. Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s415XX-XXX-XXXX-X

The COVID-19 Genomics UK (COG-UK) Consortium

William T. Harvey[1,2], Alessandro M. Carabelli[3], Ben Jackson[4], Ravindra K. Gupta[5], Emma C. Thomson[6,7], Ewan M. Harrison[3,7], Catherine Ludden[3], Andrew Rambaut[4], Sharon J. Peacock[3] and David L. Robertson[2]

[1]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
[2]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK
[3]Department of Medicine, University of Cambridge, Cambridge, UK
[4]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
[5]Cambridge Institute of Therapeutic Immunology and Infectious Disease (CITIID), University of Cambridge, Cambridge, UK
[6]Department of Clinical Research, London School of Hygiene and Tropical Medicine, London, UK
[7]Wellcome Sanger Institute, Hinxton, UK
A full list of members and their affiliations appears in the Supplementary Information.

Related links

The COVID-19 Genomics UK (COG-UK) Consortium: https://www.cogconsortium.uk

Global Report Investigating Novel Coronavirus Haplotypes: https://cov-lineages.org/global_report.html

CoV-GLUE — A Web Application for Tracking SARS-CoV-2 Genomic Variation: http://cov-glue.cvr.gla.ac.uk/

COG-UK Mutation Explorer: http://sars2.cvr.gla.ac.uk/cog-uk/

GISAID Initiative: https://www.gisaid.org

Display items

Figure 1. Spike protein sequence variability and structure. a| Schematic representation of the domain organisation of the SARS-CoV-2 spike protein showing amino acid sequence variability. The spike protein is synthesised as a 1,273 amino acid polypeptide and the frequency of amino acid variants, including both substitutions and deletions, at each of these positions is shown. These variants, relative to the reference Wuhan-Hu-1 sequence, were identified using CoV-GLUE[94], which filters out GISAID sequences[95] identified as being of low quality or from non-human hosts (sequences retrieved from the GISAID database on 3 February 2021). Among 426,623 genomes after filtering, 5,106 different amino acid replacements or substitutions across 1,267 spike positions were identified, of which 320 at 259 positions were observed in at least 100 sequences. In addition to substitutions, several deletions have been observed, particularly within the NTD. The most frequently detected amino terminal domain (NTD) deletion is the two-residue deletion at positions 69 and 70 (Δ69–70), present in 45,898 sequences. **b**| Depiction of the spike protein monomer displaying an upright receptor-binding domain (RBD). **c**| Spike protein structure in the closed conformation overlaid with surface representations shown with a trimer axis vertical view (centre) and an orthogonal top-down view along this axis (right). Domains are coloured as in part a. RCSB Protein Data Bank (PDB) IDs for the SARS-CoV-2 spike protein structures: 6ZGG and 6ZGE (ref.[50]).

NTD, amino-terminal domain; RBD, receptor-binding domain; FP, fusion peptide; CD, connecting domain; TM, transmembrane domain; C,T cytoplasmic tail; scissors, S1–S2 boundary at amino acid position 685; magenta spheres, glycans; magenta triangles, potential *N*-linked glycosylation sites.

Figure 2. Neutralising antibody classes defined by structural analyses and properties of spike protein residues. a| Amino acid residues of the SARS-CoV-2 spike protein are coloured according to the class of antibody that binds to an epitope. Receptor-binding domain (RBD) antibody classes 1–4 (ref.[31]) are shown: green (class 1: ACE2-blocking antibodies that bind the spike protein in the open conformation), yellow (class 2: ACE2-blocking antibodies that bind RBD in both open and closed conformations), blue (class 3: antibodies that do not block ACE2 and bind the RBD both in open and closed conformations) or red (class 4: neutralising antibodies that bind outside the ACE2 site and only in the open conformation). When residues belong to epitopes of multiple classes, priority colouring is given to antibodies that are ACE2 blocking and bind the closed spike. The amino-terminal domain (NTD) supersite[30] is coloured in magenta. **b**| Aligned heatmaps showing properties of amino acid residues where substitutions affect binding by antibodies in polyclonal human blood plasma or emerge as antibody escape mutations. Distance in Å to the ACE2-contacting residues that form the receptor-binding site

(RBS) is shown in shades of orange; each residue is classified as having evidence for mutations either affecting neutralisation by monoclonal antibodies (mAbs)[40,43,47,48] or polyclonal antibodies in plasma from previously infected individuals (convalescent) [39-41,43,48], or vaccinated individuals[59] (mAb effect and plasma effect, respectively). A subset of these residues have mutations described as emerging upon exposure (co-incubation) to mAbs[40,47,48] or plasma[40,41] in laboratory experiments (mAb emerge and plasma emerge, respectively). When an observation includes a deletion, this is indicated by a red cross. Shades of green depict the results of deep mutational scanning results (DMS) experiments where yeast cells expressing RBD mutants are incubated with to multiple samples of human convalescent[39]. The escape fraction (that is, a quantitative measure of the extent to which a mutation reduced polyclonal antibody binding) averaged across all amino acid substitutions at a residue (plasma average) and the maximally resistant substitution (plasma max) are indicated. DMS data on ACE2 binding affinity[19] is shown in shades of red or blue representing higher and lower ACE2 affinity, respectively. The mean change in binding affinity averaged across all mutations at each site (binding average) and alternatively the maximally binding mutant (binding max) are represented. Scores represent binding constants (Δlog10 K_D (log10)) relative to the wild-type reference amino acid.

Figure 3. Structure-based analysis of conformational epitopes on the spike protein. a| Structure-based antibody accessibility score for each spike protein ectodomain residue in the closed form were calculated using BEpro[49]. Black diamonds at the top and bottom of the plot indicate the positions of ACE2-contacting residues. Accessible amino-terminal domain (NTD) loops N1–N5 are labelled and a loop falling between these is indicated with an asterisk. **b**| Two surface colour representations of antibody accessibility scores for the spike protein in the closed conformation according to the colour scheme of part a; a trimer axis vertical view (left) and an orthogonal top-down view along this axis (right). **c**| The extent to which each spike residue becomes more or less accessible when the spike is in its open form is shown. For each spike monomer (upright RBD (yellow), closed RBD clockwise adjacent (green) and closed RBD anti-clockwise adjacent (blue), the difference relative to the score calculated for the closed form (shown in part a) is shown. **d**| Two surface colour representations of antibody accessibility scores for the spike protein in the open conformation with a single monomer with an upright receptor-binding domain (RBD) are shown; a trimer axis vertical view (left) and an orthogonal top-down view along this axis (right) are shown.

Figure 4. Structural context of spike amino acid mutations in the global virus population. Spike amino acid residues are coloured according to the frequency of amino acid substitutions or deletions. Variants (retrieved from CoV-GLUE) are based on 426,623 high quality sequences downloaded from GISAID on February 3 2021. **a**| Points representing each spike amino acid residue are positioned according to antibody accessibility score and distance to the nearest residue in the receptor-binding site. Residues with at least 100 sequences possessing a substitution or deletion are coloured according to the frequency scale shown, with the remainder shaded grey. **b**| Spike in closed form with all residues coloured according to the frequency scale shown; a trimer axis vertical view (left) and an orthogonal top-down view along this axis (right) are shown. **c|** A close-up view of the receptor-binding domain (RBD) bound to ACE2 (RCSB Protein Data Bank ID: 6M0J (ref.[96])), with RBD residues shown as spheres coloured by amino acid variant frequency and ACE2 shown in gold. Amino acid variants are present at high frequency in positions at the RBD–ACE2 interface. (**d**) Spike in open form with residues where at least 100 sequences possessing a substitution are highlighted; a trimer axis vertical view (left) and an orthogonal top-down view along this axis (right) are shown.

Figure 5. Amino acid mutations that characterise variants of concern. a| Spike heterotrimer in the open conformation overlaid with surface representation (RCSB Protein Data Bank ID: 6ZGG (ref.[50])). Locations of amino acid substitutions and deletions that define variants of concern are highlighted as red spheres. For B.1.1.7, scissors mark the approximate position of substitution P681H within the furin cleavage site which is absent from the structural model. **b**| Aligned heatmaps showing properties of amino acid residues or of the specific amino acid substitution, as appropriate. Epitope residues are coloured to indicate amino-terminal domain (NTD) or receptor-binding domain (RBD) class[30]. Structure-based antibody access scores for the spike protein in closed

and open conformations are shown. For RBD residues, the results of deep mutational scanning (DMS) studies show the escape fraction (that is, a quantitative measure of the extent to which a mutation reduced polyclonal antibody binding) for each mutant averaged across plasma (average) and for the most sensitive plasma (plasma max)[39]. Each mutation is classified as having evidence for mutations either affecting neutralisation by monoclonal antibodies (mAbs) or antibodies in convalescent plasma[39], or vaccinated individuals [59], and emerging in selection experiments using mAbs[40,47,48] or post-infection serum[40,47,48]. Distance to the ACE2-contacting residues that form the receptor-binding site RBS (Å) is shown (for residue 681, this is estimated using the nearest residues present in published structures). DMS data on ACE2 binding affinity[19] is shown by aggregating scores and averaging across each mutant at a residue and alternatively the maximally binding mutant. Scores represent binding constant ($\Delta\log_{10} K_D$) relative to the wild-type reference amino acid. Mutations that are present in a variant but that are also widespread in the virus population in which a variant emerged are marked with †.

Box 1 | Spike protein structure and function

As with other coronaviruses, the entry of SARS-CoV-2 into host cells is mediated by the transmembrane spike glycoprotein, which forms homotrimers on the surface of the virion. The SARS-CoV-2 spike protein is highly glycosylated, with 66 potential *N*-glycosylation sites per trimer (**Figure 1a**)[97,98].The SARS-CoV-2 spike is post-translationally cleaved by mammalian furin into two subunits: S1 and S2 (**Figure 1a**). The S1 subunit largely consists of the amino-terminal domain (NTD) and the receptor-binding domain (RBD), and is responsible for binding to the host cell surface receptor, angiotensin-converting enzyme 2 (ACE2), whereas the S2 subunit includes the trimeric core of the protein and is responsible for membrane fusion (**Figure 1b**). The presence of a polybasic furin cleavage site at the S1–S2 boundary, unique within the *Sarbecovirus* subgenus, is important for infectivity and virulence[99], with furin cleavage facilitating the conformational change required for receptor-binding[50]. The spike protein transiently undergoes conformational changes between a closed and an open conformation in which a hinge-like movement raises the RBD[50]. The residues comprising the receptor-binding motif (RBM) are revealed on the upright RBD, enabling binding to ACE2, which induces a progressively more open structure until a fully open, three-ACE2-bound structure is formed, initiating S2 unsheathing and membrane fusion[100].

Box 2| Mechanisms of antigenic change

In common with other virus surface glycoproteins responsible for attachment to host cell surface receptors, such as influenza virus haemagglutinin and the envelope glycoprotein GP120 of HIV, the SARS-CoV-2 spike glycoprotein is an important target for neutralising antibodies. There are various distinct mechanisms by which mutations can alter the antigenic properties of a glycoprotein.

Amino acid substitutions that alter the epitope

A change in the biophysical properties of an epitope residue directly diminishes antibody binding. For example, the neutralising antibody 4A8 forms salt bridges with spike protein residues K147 and K150, and therefore substitutions to these residues are likely to inhibit binding. The E484K amino acid substitution has received attention for its effect on mAbs and convalescent plasma neutralising activity. Its position has been described as belonging to the footprint of several antibodies and a change in charge caused by a replacement of a glutamic acid with a lysine has the potential to diminish antibody binding.

Increasing receptor-binding avidity

Substitutions that individually increase receptor-binding affinity can shift the binding equilibrium between glycoprotein and neutralising antibodies in favour of a higher avidity interaction between

glycoprotein and cellular receptor [101]. The spike amino acid substitution N501Y, which increases ACE2-binding affinity[19], has been described as emerging in individuals treated with convalescent plasma, potentially as a means of immune-escape.

Changes in glycosylation

Glycans are bulky sugar molecules that may shield epitopes from antibody binding. *N*-linked glycans are typically prominent in glycan shielding of virus surface glycoprotein epitopes [33], though *O*-linked glycans can also contribute[102]. A substitution can introduce an additional *N*-linked glycosylation motif. The acquisition of epitope-masking glycans during the evolution of human influenza viruses is well described [103].

Deletions and insertions

The deletion or insertion of residues has the potential to alter epitope conformation diminishing antibody binding. Several deletions in the spike NTD that affect recognition by neutralising antibodies have been described[41,42]. In laboratory experiments, a multi-residue insertion in the spike NTD has been described as emerging and contributing to escape from polyclonal antibodies in convalescent plasma[41].

Allosteric structural effects

Similar to deletions or insertions, an amino acid substitution outside of an epitope footprint may affect antibody binding by changing the protein conformation in such a way that an epitope is altered or differently displayed. In the spike NTD, changes to disulfide bonds are thought to reduce binding by multiple mAbs through this mechanism[30].

Glossary

Mutation. The substitution, insertion or deletion of one or more nucleotides in the virus RNA genome. Non-synonymous nucleotide substitutions in protein-coding sequence result in a change in amino acid (referred to as a substitution or replacement), whereas synonymous nucleotide substitutions do not change the amino acid.

Amino acid substitution. By convention, an amino acid substitution is written N501Y to denote the wild-type (N (asparagine)) and substituted amino acid (Y (tyrosine)) at site 501 in the amino acid sequence.

dN/dS ratio. The ratio of non-synonymous mutations per non-synonymous sites (dN) to synonymous mutations per synonymous sites (dS) is used to estimate the balance between neutral mutations, purifying selection and positive selection acting on gene or a specific codon.

Glycoprotein. A protein with oligosaccharide chains (glycans) covalently attached to amino acid side chains. Virus surface glycoproteins embedded in the membrane often have a role in interactions with host cells including receptor-binding and are also commonly targeted by host antibodies.

Epitope. The specific part of an antigen recognised by the immune system, either antibodies, B-cells or T-cells.

Epitope mapping. Experimental determination of the binding site, or epitope, of an antibody. Approaches include X-ray co-crystallography or cryogenic electron microscopy of an antigen-antibody complex and the mapping of systematic mutations introduced by site-directed mutagenesis.

Antibody footprint. Antibodies often recognise conformational epitopes that consist of amino acid residues brought together by the 3D folding of a protein. Residues on the part of the protein contacted by a binding antibody form its footprint.

Glycan-shielding. Glycans are bulky sugar molecules covalently attached to amino acid side chains that can cloak areas of the underlying protein impeding antibody binding.

Monoclonal antibody (mAb). An antibody made by cloning a unique white blood cell, which usually has monovalent binding affinity for a specific epitope. Virus particles can be saturated with mAbs and the structure solved to determine the antibody footprint or mAbs can also be used to select for mutations that escape recognition.

Epitope binning. A competitive immunoassay is used to sort a library of monoclonal antibodies targeting a protein to determine, which compete for access to overlapping epitopes that are then binned together into antigenic sites.

Convalescent plasma. Blood serum of a previously infected individual that usually contains a mixture of different antibodies referred to as polyclonal antibodies. Similarly, post-vaccination serum includes polyclonal antibodies generated by vaccination.

Immunodominance. Differing levels of antigen immunogenicity results in a dominance hierarchy such that the immune response is mostly focussed on a few antigens, with more potently neutralising antibodies targeting immunodominant epitopes.

Viral variant. A genetically distinct virus with different mutations to other viruses. Variants can also refer to the founding virus of a cluster or lineage and used to refer collectively to the resulting variants that form the lineage. Variants with changed biological characteristics or antigenicity have been termed variants of interest or concern by public health bodies.

Lineages. Monophyletic clusters of viruses assigned based on the SARS-CoV-2 global phylogenetic tree.

IC_{50}. The half maximal inhibitory concentration is a quantitative measure that indicates how much of an inhibitory substance, e.g., post-vaccination serum, is required to inhibit a biological process, e.g., virus forming plaques or regions of infected cells in culture, by 50%.

Avidity. Also referred to as functional affinity, avidity is the accumulated binding strength of multiple affinities of individual interactions such as between a virus receptor binding site and its cellular receptor.

Table of contents blurb

SARS-CoV-2 evolution has been characterised by the emergence of mutations and so-called ‘variants of concern’ that impact virus characteristics including transmissibility and antigenicity. In this Review, members of the COVID-19 Genomics UK (COG-UK) Consortium and colleagues summarise mutations to the SARS-CoV-2 spike, focussing on their impacts on antigenicity and contextualising them in the protein structure, and observed mutation frequencies in global sequence datasets.