COG-UK update on SARS-CoV-2 Spike mutations of special interest

Report 1

Prepared by COG-UK, 19th December 2020

Summary

This report provides **background context on mutation tracking by COG-UK** and describes a **priority set of SARS-CoV-2 Spike mutations** that are of particular interest based on potential epidemiological significance in the UK and/or biological evidence based on the literature or unpublished work. It provides details on the frequency of mutations, and their potential biological and immunological significance as we currently understand it. At this point in time, there is no reason to believe that any of the mutations discussed here will affect vaccine efficacy. **Appendix 1** provides explanations for terms (mutation, variant, lineage) and the basis for prioritising the mutations described here.

The analysis described below is based on 126,219 genomes from positive samples generated by the COG-UK consortium. This identified 1,777 different **amino acid changing** (non-synonymous) mutations in Spike glycoprotein's gene S (this does not include mutations that do not lead to an alternation of amino acid (which are more numerous), or mutations elsewhere in the genome). Of these non-synonymous changes, 37% (n=654) mutations were only observed in a single sequence, while 5% (n=87) were observed in at least 100 sequences.

Five amino acid replacements (D614G, A222V, N439K, Y453F and N501Y), one deletion (del) and co-occurrence of some of these changes are actively being investigated by COG-UK. Further details, including the reason for their inclusion, is provided in **Table 1**. This is a shortlist of Spike-focused priority mutations, but others are being monitored in S and other SARS-CoV-2 genes.

The lineage B.1.1.7[1] is of particular interest (see footnote), and is notable for a higher number of mutations in one lineage than observed previously (**Table 1b**). It has been speculated that it may have arisen from a chronically infected individual. One of these (the N501Y mutation) occurs in the region of the Spike protein, the receptor binding domain (RBD), that the virus uses to bind to the human ACE2 receptor. Changes in this region of the Spike protein can result in the virus changing its ACE2 binding specificity and alter antibody recognition.

Two other mutations (N439K and Y453F) also occur in the RBD region and increase binding affinity to ACE2, and have been shown to escape the neutralising effect of a few monoclonal antibodies (mAbs). The 69-70del has co-occurred with all three of these RBD mutations.

Limitations

COG-UK undertakes sequencing of SARS-CoV-2 samples from about 10% of people from across the UK who are confirmed positive cases. This in turn is an underestimate of the true number of infections. The sequenced genomes are used to identify mutations that arise and to identify which of the viral proteins they could affect.

[1] Named by Public Health England as VUI-202012/01 (the first "Variant Under Investigation" in December 2020).

Distribution of specific mutations in the UK

The cumulative number of instances that we have detected these mutations of interest in the UK are shown in **Table 1a**. As described in limitations above, this will be an under-representation, but provides insights into proportions in the dataset.

D614G was not present when the virus first emerged but is now ubiquitous and appears to be associated with a moderate effect on SARS-CoV-2 transmissibility.

A222V is present in the 20A.EU1 SARS-CoV-2 'cluster' (also designated as lineage B.1.177), which has been spreading in Europe and seems to have originated in Spain. Multiple introductions have occurred into the UK followed by transmission across the country, suggesting that this spread was likely associated with travel to/from Spain over the summer.

N439K first emerged in Scotland in March 2020 in a lineage that is no longer circulating (B.1.141), but has since been introduced into the UK on multiple occasions (in lineage B.1.258). This is now circulating in many European countries and internationally.

N501Y and 69-70del are present in B.1.1.7, which has been growing in frequency since November 2020 and is defined by a set of 14 amino acid changes and 3 deletions. This is responsible for an increasing proportion of SARS-CoV-2 cases in the UK. Further updates on frequency are being prepared by COG-UK and Public Health Agencies. N501Y is also present in a lineage of independent origins circulating in Wales (this lineage does not have the other mutation seen in B.1.1.7).

Y453F has received widespread attention because it has been observed in the context of mink-human infections and it appeared in the widely reported Danish Mink cluster but has not been observed in the UK to date.

Table 1b provides information on the priority lineages being tracked by COG-UK.

COG UK
COVID-19 GENOMICS UK CONSORTIUM

Table 1a. Priority mutations being tracked by COG-UK

Mutation	Predominant Lineage	Reasons for tracking	Cumulative number in UK	Number over last 28 days (13/11/2020 - 10/12/2020)
D614G	B.1	Moderate effect on transmissibility	118,906	11,447
A222V	B1.177	Fast growing lineage but no evidence of mutation effect	46,710	7,856
N439K	B.1.141 B.1.258	1) Increased binding affinity to hACE2 receptor 2) Escape to some mAbs	3,320	246
Δ69-70	B.1.1 B.1.258	1) Evasion immune response 2)Diagnostic failure in some assays targeting the S gene including the three-target TaqPath assay and the two-target Biofire assay	3,504	1,228
N501Y	B.1.1.7	Fast growing lineage & increased binding affinity to hACE2 receptor	2,057	1,182
N501Y + Δ69-70	B.1.1.7	Likely to maintain characteristics described for N501Y and 69-70del	1,524	1,034
N439K + Δ69-70	B.1.258	Likely to maintain characteristics described for N439Y and 69-70del	1,895	176
Y453F	B.1.1 B.1.1.298	1) Increase binding affinity to hACE2 receptor 2) Escape to some mAbs Human/mink associated	0	0

Cumulative number as of December 15 based on data deposited into CLIMB. Caution is required since the data will not include information from the last 2 weeks.

Table 1b. Priority lineages being tracked by COG-UK

variant	Reason for tracking	Cumulative number in the UK	Number over last 28 days (13/11/2020 - 10/12/2020)
'Cluster 5 variant'	Danish Mink variant. Contains 4 mutations including: Y453F, 69-70del, I692V and M1229I . Cluster 5 variants may be able to escape the effect of convalescent plasma. Y453F has increased binding affinity to the human ACE2 receptor in laboratory experiments	0	0
B.1.1.7 (variant)[1]	Has 17 mutations (14 replacements and 3 deletions) including: T1001I, A1708D, I2230T, SGF 3675-3677 del In the ORF1ab; 69-70 del, Y144 del, N501Y, A570D, P681H, T716I, S982A and D1118H in the Spike; Q27stop, R52I and Y73C in ORF8; D3L and S235F in the N. Noteworthy N501Y enhances ACE2 binding affinity, 69-70del has immunological role and it is associated with some diagnostics failures, and P681H occurs at the furin cleavage site, known for biological significance in membrane fusion	1416	945

[1]Named by Public Health England as VUI-202012/01 (the first "Variant Under Investigation" in December 2020)

Potential biological significance of mutations

B.1.1.7 lineage. This variant has 17 mutations (the 14 amino acid replacements and 3 in-frame deletions are listed in **Table 1b**). Two of these mutations have already been described to alter SARS-CoV-2 biology: N501Y sits in the receptor binding motif (RBM) of the Spike protein, and has been described to increase binding affinity to the human ACE-2 receptor; 69-70del has been identified in variants associated with immune escape in immunocompromised patients and is responsible for a "dropout" in the S gene PCR target in certain diagnostic tests (e.g. Thermo Fisher TaqPath). These tests target multiple regions of the virus genome, so the test itself is not compromised. Reported cases and phylogenetic analyses have indicated an exceptional rate of introduction of mutations into this lineage. It has been hypothesised that this lineage may have resulted from the transmission of the virus from a chronically-infected individual. This is based on observations that a high rate of mutations may accumulate In immunocompromised patients with chronic infections of SARS-COV-2.

N439K. There is no evidence for a faster rate of growth for the 439K variant beyond that already determined for the D614G mutation which is also found in all variants carrying 439K. N439K enhances

binding affinity to the hACE2 receptor and is able to escape the neutralising activity of some mAbs, including one in clinical trials, and from some antibodies present in sera from a sizable fraction of people recovered from infection. By looking at the medical interventions needed/outcomes for patients carrying either 439N or 439K, no increased disease severity was observed. Furthermore, there is no evidence that this mutation will allow the virus to escape immunity triggered by vaccines.

Y453F was identified in the Netherlands and Denmark associated with mink-human infections and in particular in Danish cluster 5 which was defined as a cluster of variants carrying four mutations, the 69-70del and three amino acid replacements (Y453F, I692V and M1229I) in the Spike protein, and spreading among farmed mink and a small number of people in Denmark (**Table 1b**). This lineage raised concerns because of a reported reduction of neutralization activity of sera from people recovered from infection, but further studies are required. The Y453F mutation, which occurs in the RBM, although rare, has also been observed previously in SARS-CoV-2 genomes isolated from humans, but never in the UK. This mutation has arisen independently multiple times in several countries. Y453F has also been shown in laboratory studies to increase the affinity of Spike protein binding to the hACE2 receptor.

A222V. It has been speculated that the increased transmissibility of the B.1.177 lineage may be associated with the presence of the A222V mutation, but more evidence is needed.

Structural context for mutations can be found in Appendix 2.

Key COG-UK references

COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk. 2020. “An Integrated National Scale SARS-CoV-2 Genomic Surveillance Network.” The Lancet. Microbe 1 (3): e99–100.

Bal, A., Destras, G., Gaymard, A., Regue, H., Semanas, Q., d’Aubarde, C., Billaud, G., Laurent, F., Gonzales, C., Valette, M., Lina, B., Morfin, F., Josset, L. “Screening of the H69 and V70 deletions in the SARS-CoV-2 spike protein with a RT-PCR diagnosis assay reveals low prevalence in Lyon, France.” medRxiv 2020, 2020.11.10.20228528.

Kemp, Steven A., Dami A. Collier, Rawlings Datir, Salma Gayed, Aminu Jahun, Myra Hosmillo, Isabella Atm Ferreira, et al. 2020. “Neutralising Antibodies Drive Spike Mediated SARS-CoV-2 Evasion.” medRxiv. https://www.medrxiv.org/content/10.1101/2020.12.05.20241927v1.full.

Rambaut, Andrew, Edward C. Holmes, Áine O’Toole, Verity Hill, John T. McCrone, Christopher Ruis, Louis du Plessis, and Oliver G. Pybus. 2020. “A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology.” Nature Microbiology 5 (11): 1403–7.

Thomson, E. C., L. E. Rosen, J. G. Shepherd, and R. Spreafico. 2020. “The Circulating SARS-CoV-2 Spike Variant N439K Maintains Fitness While Evading Antibody-Mediated Immunity.” bioRxiv. https://www.biorxiv.org/content/10.1101/2020.11.04.355842v1.abstract.

Volz, Erik, Verity Hill, John T. McCrone, Anna Price, David Jorgensen, Áine O’Toole, Joel Southgate, et al. 2020. “Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity.” Cell, November. https://doi.org/10.1016/j.cell.2020.11.020.

Appendix 1

Background

Mutations arise naturally in the SARS-CoV-2 genome as the virus replicates and circulates in the human population. As a result of this on-going process, many thousands of mutations have already arisen in the SARS-CoV-2 genome since the virus emerged in late 2019. As mutations continue to arise, novel combinations of mutations are increasingly observed. The vast majority of mutations have no apparent effect on the virus. Only a very small minority are likely to be important and change the virus in any appreciable way. This could include a change in the ability to infect/transmit between people; a change in disease severity; or a change in the way the virus interacts with the immune system (including the response generated by a vaccine). We pay most attention to mutations in the gene that encodes the Spike protein, which is associated with viral entry into cells and it is relevant in the context of immunity and vaccine efficacy.

Definitions

Mutation is used to describe a change of a nucleotide in the virus RNA genome, a subset of which results in a change in amino acid (sometimes referred to as a substitution or replacement), or a mutation can refer to a deletion or insertion event in the virus genome. By convention an amino change is written N501Y to denote the wildtype (N, asparagine) and replacement amino acid (Y, tyrosine) at site 501 in the amino acid sequence.

Viral variant refers to a distinct virus, which may have a combination of different mutations and may appear independently. Variant refers to the founding genetic sequence and the resulting genomes form a lineage (they will then diverge from the variant).

Lineages are assigned combining genetic and epidemiological data. Due to the naturally expanding genetic diversity of SARS-CoV-2 a nomenclature system has been introduced by Rambaut et al. (2020), see https://cov-lineages.org.

How do we choose which mutations or variants to track?

It is difficult to predict whether any given mutation is important when it first emerges, against a backdrop of the continuous emergence of new mutations. We take several approaches to identify mutations of interest:

1. We look for mutations of theoretical concern, that is, these have been identified as potentially important in laboratory experiments but have not arisen in people yet but if they did, we would want to rapidly understand if this was important in humans.
2. We look for trends in the frequency of specific viral variants. If these are appearing more often in the population than other variants, there are several explanations that need to be discounted. But one possible reason is that a virus with a specific mutation or combination of mutations may spread more rapidly in the population as a result of increased infectivity or transmissibility.
3. We respond to the Public Health Agency concern that there appears to be a group of people who may have more severe cases of infection, supporting the sequencing and analysis of their genomes.
4. As vaccines are rolled out, it will be important to sequence SARS-CoV-2 virus from infected people who have been vaccinated, or have had a second episode of COVID-19. The aim is to detect variants that are evading the immune system elicited by past infection or vaccination.

Appendix 2

The extent to which SARS-CoV-2 may evolve to escape immunity induced by infection or vaccination is not currently known. Determining phenotype from genetic data is a fundamental challenge. **Figure 1a** shows the localisation of the selected mutations in a three dimensional structure of the Spike protein. A222V and the 69-70 deletion are localised relatively far from the receptor-binding site in comparison with amino acid residues 453, 439 and 501 which are in the RBD region. For each amino acid present in the Spike structure, an antibody accessibility score was calculated in **Figure 1b**. High antibody accessibility scores for 501, 439 and 70 correspond to sites that sit on the surface of the protein and that are more easily accessible to antibodies. Antibodies (Ab) are known to recognise specific regions of the Spike protein known as epitopes. Depending on the areas that Abs target there are 4 classes for the RBD region and 1 class for the N-terminal domain (NTD) near to where 69-70 sit (**Figure 1c-d**).

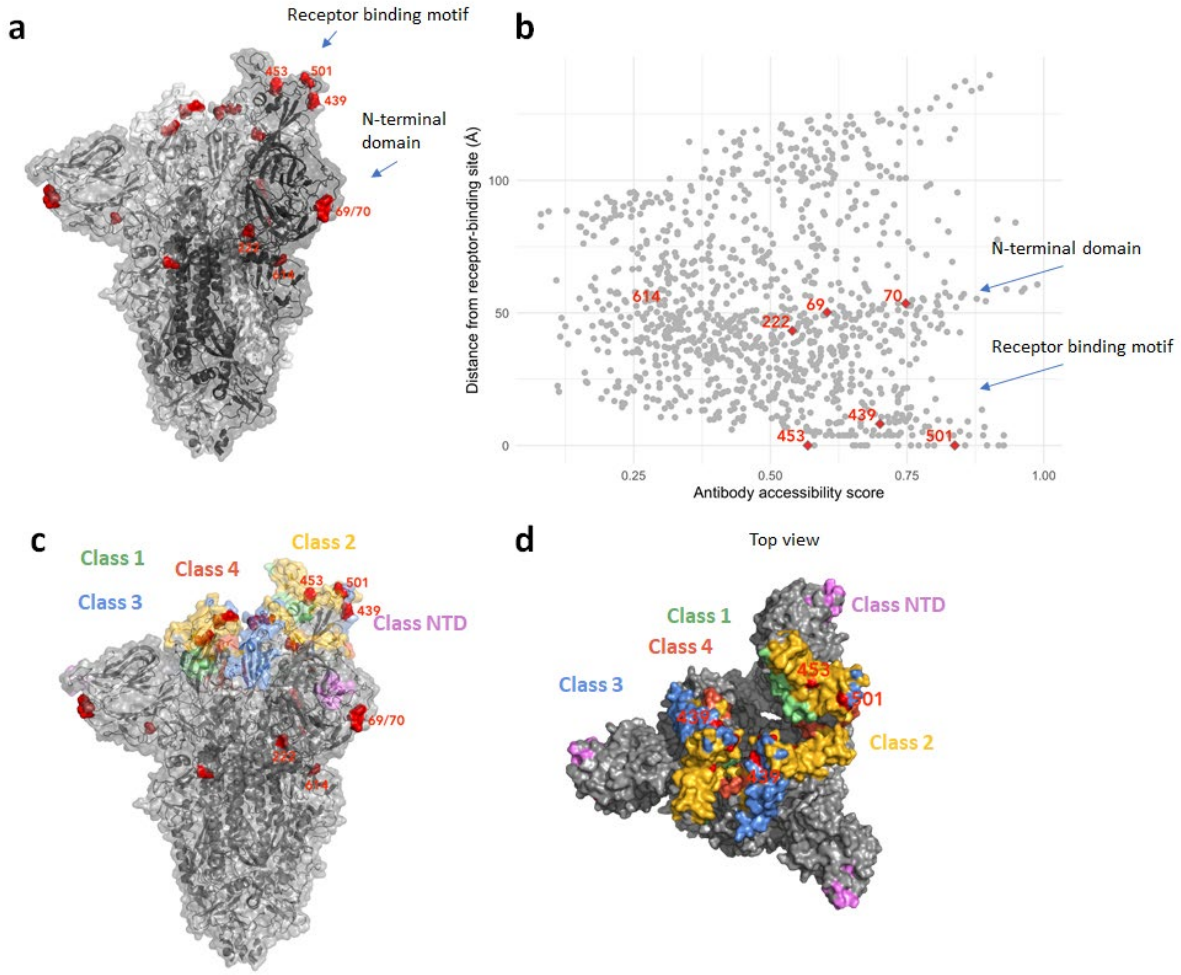

Figure 1. Localisation of mutations in the Spike structure. a) Spike heterotrimer in open conformation (PDB: 6ZGG, Wrobel et al. 2020). Locations of deleted residues His69 and Val70 and the residues involved in substitutions (A222V, N439K, Y453F, and N501Y) are highlighted in red; **b)** Each point represents a Spike protein amino acid residue positioned according to distance from the hACE2 receptor-binding site and an antibody accessibility score. Residues

associated with high interest amino acid substitution or deletions are highlighted with red diamonds. Residues belonging to the receptor-binding site defined as those with atoms within 4Å in Spike:hACE2 complex and distance to these residues based on closed conformation Spike. Antibody accessibility score represents surface accessibility and amino acid identity of target residue and weighted average of nearby residues and is scaled between minimum 0 and maximum 1, calculated across Spike in open and closed conformations; residues are positioned according to their maximum score across Spike in either open and closed conformations; **c-d)** Highlighted in colours regions target by different classes of Abs. 453, 501 and 439 are localised in the regions targeted by some classes of mAbs. 69-70 is near a region targeted by other mAbs and deletion might alter the structure of neighbouring amino acids. green = class 1: ACE2 blocking, bind open RBD only; yellow = class 2: ACE2 blocking, bind open and closed RBD; blue = class 3: non-ACE2 blocking, bind open and closed RBD; orange = class 4: non_ACE2 blocking, bind open RBD only). Epitope residues described in the NTD are coloured in magenta.