Recurrent SARS-CoV-2 Mutations in Immunodeficient Patients

Sam AJ Wilkinson¹, Alex Richter², Anna Casey¹, Husam Osman³, Jeremy D Mirza¹, Joanne Stockton¹, Josh Quick¹, Liz Ratcliffe³, Natalie Sparks¹, Nicola Cumley¹, Radoslaw Poplawski¹, Sam Nicholls¹, Beatrix Kele⁴, Kathryn Harris⁵, The COVID-19 Genomics UK (COG-UK) consortium⁶, Thomas P Peacock⁵*, Nicholas J Loman¹*

¹: Institute of Microbiology and Infection, School of Biosciences, University of Birmingham, UK, B15 2TT
²: Institute of Immunology and Immunotherapy (III), College of Medical and Dental Sciences, University of Birmingham, UK, B15 2TT
³: Queen Elizabeth Hospital, University Hospitals Birmingham, Birmingham, UK, B15 2TH.
⁴: Virology Department, NHS East and South East London Pathology Partnership, Royal London Hospital, Barts Health NHS Trust, UK, EC1A 7BE.
⁵: https://www.cogconsortium.uk - Full list of consortium names and affiliations are in the appendix
⁶: Department of Infectious Disease, Imperial College London, London, UK, W2 1PG

*: Corresponding Authors

Abstract

Long-term SARS-CoV-2 infections in immunodeficient patients are an important source of variation for the virus but are understudied. Many case studies have been published which describe one or a small number of long-term infected individuals but no study has combined these sequences into a cohesive dataset. This work aims to rectify this and study the genomics of this patient group through a combination of literature searches as well as identifying new case series directly from the COG-UK dataset. The spike gene receptor binding domain (RBD) and N-terminal domains (NTD) were identified as mutation hotspots. Numerous mutations associated with variants of concern were observed to emerge recurrently. Additionally a mutation in the envelope gene, - T30I was determined to be the most recurrent frequently occurring mutation arising in persistent infections. A high proportion of recurrent mutations in immunodeficient individuals are associated with ACE2 affinity, immune escape, or viral packaging optimisation.

There is an apparent selective pressure for mutations which aid intra-host transmission or persistence which are often different to mutations which aid inter-host transmission, although the fact that multiple recurrent de novo mutations are considered defining for variants of concern strongly indicates that this potential source of novel variants should not be discounted.
Introduction

Long-term SARS-CoV-2 infections in immunodeficient patients are important, but understudied (Moran et al., 2021). Evolution of viruses during long-term infection is an important source of novel variation and is thought to be a key influence of the evolutionary dynamics of SARS-CoV-2 generally, and the emergence of new variants specifically. Notably Alpha and Omicron, which were responsible for recent epidemic waves globally, are hypothesised by some to have arisen during long-term infections (Msomi et al., 2021; Rambaut et al., 2020).

The Alpha variant (B.1.1.7) emerged abruptly with a constellation of novel mutations and a long branch length from its nearest common ancestor in the B.1.1 clade, during a time of extremely high surveillance in the UK (Rambaut et al., 2020). A likely explanation is that the Alpha variant evolved within a single long-term host over a long period before emergence back into the general population. Evolution during long-term infection has been associated with the rapid accumulation of many mutations within a short period (Avanzato et al., 2020; Baang et al., 2021; Choi et al., 2020; Jensen et al., 2021; Karim et al., 2021; Peacock et al., 2021; Riddell et al., 2022). The Beta (B.1.351), Gamma (P.1), and Omicron (B.1.1.529) variants all emerged in similar circumstances to alpha, potentially suggesting that they also emerged from long-term infections.

To better understand evolutionary pressures associated with viral evolution during long-term infections a dataset composed of 168 SARS-CoV-2 genomes associated with 28 patients with a range of conditions that result in immunodeficiency significant enough to prevent rapid viral clearance was compiled to examine the frequency of recurrent mutations. This builds upon previous work performing a similar analysis using case studies which included a total of ten patients (Peacock et al., 2021). This analysis expands on previous studies by utilising a significantly larger dataset which increases the power, also many of the cases included are the alpha variant which have not been discussed in the context of long-term SARS-CoV-2 cases previously and potentially gives insight into future variant emergence, and lastly all genome series’ were analysed using a single analysis pipeline.
Methods

Dataset assembly

Patient-associated genome series were selected for inclusion via a literature search for case studies using the following search terms and filters: After 2019, “SARS-CoV-2”, “nCoV-2019”, “Immunodeficient”, “Immunocompromised”, “long-term”, all searches took place between the dates 01/08/2021 and 30/11/2021. Other genome series were extracted from the COG-UK dataset, a UK-wide genomic surveillance repository (COVID-19 Genomics UK (COG-UK), 2020; Nicholls et al., 2021).

Genome series were only included if they met the following criteria: at least two genomes available on either public databases or via a request, evidence of long-term viral infection for a period no less than 28 days (some genome series covered a shorter period but the clinical information met this criterion), clinical information available was sufficient to indicate the nature of the patient's immune deficiency. For all genome series included in the dataset a Civet report (O'Toole et al., 2021a) was generated using Civet v3.0. These reports confirm that all genomes were the result of long-term infections rather than a super-infection or independent infection events by virtue of individual genomes sharing a recent common ancestor with a step-wise accumulation of mutations over time. A single genome from patient 11 was excluded due to a probable superinfection as described by (Tarhini et al., 2021). Figures were generated for each phylogeny generated with civet using ggtree (Yu et al., 2018) and are included within the supplementary material.

When a genome series was selected for inclusion all genomes were placed within an individual multi-fasta file with a header identifying the patient via an identifier (“pt-1”, pt-2” etc) and the number of days passed since the initial genome available within that genome series (the day 0 genome), in several cases this genome was collected after a lengthy period of active infection but only the time period covered by the genome series was considered in the analysis.
Mutation Calling of Genomes

Mutation calling was automated with an R script adapted from (Mercatelli et al., 2021) which utilises Nucleotide mummer (NUCmer) (Marçais et al., 2018) for genome alignment to an annotated SARS-CoV-2 reference sequence (Wu et al., 2020) and defines SNPs, insertions, deletions, frameshifts, and inversions relative to this reference sequence (NCBI accession NC_045512.2). One change was made to the annotations of the reference in the case of the ORF1ab polyprotein gene NSP12 where the position was adjusted by a single nucleotide so that all mutation calls would be relative to the reading frame post the ribosomal frame-shift for simplicity; zero mutations were detected in the pre-ribosomal frame-shift region of NSP12 therefore no mutations were incorrectly annotated as a result.

De-novo Mutation Cumulative Occurrence Analysis Pipeline

Processing of the mutation calls was performed with a Python script (https://github.com/BioWilko/recurrent-sars-cov-2-mutations/blob/main/mutation_call_analysis.py) to investigate de novo mutations, which were defined as observed mutations within a genome series which were not present at day 0 of the genome series. A cumulative count of each observed de novo mutation (DNM) was performed for each day between 0 and the maximum genome series length (218 days). When a deletion was observed all deletions with a reference position within 18 nucleotides of the reference position of the initial deletion regardless of length or position were clustered as a single region. Ambiguous nucleotides were not considered in mutation calling. The resultant dataframe was finally formatted with an R script and figures generated using ggplot2 (Wickham, 2016).
Results

Figure 1: Distribution of De Novo Mutations included in this study across the entire SARS-CoV-2 genome. Schematic of SARS-CoV-2 genome with relevant ORFs annotated. DNMs with the highest frequency annotated by amino acid position and substitutions - X indicates multiple amino acids form DNMs at this position.
Figure 2: Cumulative occurrences of non-synonymous de-novo mutations in S-gene divided by gene domain in 168 genomes obtained from twenty-eight patients. Substitution mutations were clustered by amino acid loci, this is notated with the IUPAC ambiguity code X to indicate any possible amino acid. Only loci which were notable when clustered (significant difference with non-clustered equivalent, or loci not highlighted without clustering) were included in the figure. Mutations were observed in the following domains; the S1/S2 Furin Cleavage Site, Fusion Peptide (FP), N-terminal Domain (NTD), Receptor Binding Domain (RBD), Signal Peptide (SP), and Other S2 (Xia, 2021). Deletions (Δ) were clustered within a window of 6 amino acids (AA) regardless of length or position of deletion; full details of the breakdown can be found at https://github.com/BioWilko/recurrent-sars-cov-2-mutations/blob/main/dataset/mutation_calls.csv. The first genome from each patient was considered to be day 0. The sampling periods and frequencies within the dataset was highly variable, 218 days was the longest time-period covered within the dataset but the majority were much shorter, the full details of the dataset are available in supplementary table 1. All mutations with cumulative frequencies ≥ 2 were labelled on-graph.

The SARS-CoV-2 spike gene (S) demonstrated the greatest number of recurrent mutations in the dataset (Figure 2, Figure 1) with 10 substitutions - S:S13I, S:T95I, S:G142V, S:L452R, S:E484K, S:E484G, S:F486I, S:F490L, S:Q493K, and S:Q498R. The domain where the highest number of DNM occurrences were observed was the Receptor Binding Domain (RBD) with seven, followed by the N-terminal Domain (NTD) with five, and the signal peptide (SP) with one for a total of thirteen. Clustering mutations by AA loci additionally revealed the following sites as notable: S:484, S:501, S:330, and S:440. The domain with the highest number of AA loci with DNMs was the RBD with nine, followed by the NTD with five, and the SP with one. The most frequently occurring DNM was S:E484K with 8 occurrences, when all DNMs at the S:484 loci are clustered (Figure 2); the
number of occurrences is increased to twelve clearly demonstrating an enrichment of DNMs at this loci. The DNMs at the loci S:484 consist of: eight S:E484K, two S:E484G, and one each of S:E484Q, and S:E484A. AA loci clustering highlighted the loci S:330, S:440, and S:501 as recurrent for DNMs (≥ two occurrences in the period).

The only recurrent deletions observed in the dataset were located within the NTD of S-gene; S:Δ67 region (recurrent deletion region 1/RDR1), S:Δ138 region (RDR2), and S:Δ243 region (RDR4) (McCarthy et al., 2021). S:Δ138 region was the most frequent with four occurrences, followed by S:Δ67 region and S:Δ138 region with two occurrences respectively. Deletions within the S:67 region consisted of one S:Δ67 and one S:Δ69-70, the unconventional annotation is the result of the algorithm utilised to cluster deletions, the genome series in which S:Δ67 occurred already possessed S:Δ69 in its day 0 genome. S-gene constitutes just over one eighth of the overall SARS-CoV-2 genome by length; despite this ~34% (79/234) of the total DNM occurrences were observed within S-gene as well as 59% (13/22) of the recurrent DNMs.
Figure 3: Cumulative occurrences of non-synonymous de-novo mutations in genes other than S or ORF1ab subdivided by gene in 168 genomes obtained from 28 patients. Mutations were observed in the following genes; E (encodes envelope protein), M (encodes membrane glycoprotein), N (encodes nucleocapsid phosphoprotein), ORF10 (encodes ORF10 protein), ORF3a (encodes ORF3a protein), ORF6 (encodes ORF6 protein), ORF7a (encodes ORF7a protein), and ORF8 (encodes ORF8 protein) the full details of the gene definitions used are available from (Wu et al., 2020). The first genome from each patient was considered to be day 0. The sampling periods and frequencies within the dataset was highly variable, 218 days was the longest time-period covered within the dataset but the majority were much shorter, the full details of the dataset are available in supplementary table 1. All mutations with cumulative frequencies ≥ 2 were labelled on-graph.

Non-spike, non-ORF1ab SARS-CoV-2 genes demonstrated a lower number of DNM occurrences (Figure 3, Figure 1). Three mutations within Matrix (M) and Envelope (E) were notable in their frequency (≥ 2 occurrences in the period); E:T30I, and M:H125Y. E:T30I was the only recurrent DNM observed within E-gene and the second most frequent DNM revealed by the analysis overall at 6 occurrences. E:T30I occurrences were not observed to be associated with any particular source study, geographical region, or SARS-CoV-2 lineage suggesting this may be a sensitive marker for persistent infection. Within M-gene M:H125Y was the only recurrent DNM with four occurrences.
When DNMs observed in these genes were clustered by AA loci the findings remained almost entirely unchanged other than in the case of the loci M:2 which was raised to three DNM occurrences by day 218 rather than the two presented in (Figure 3).

Figure 4: Cumulative occurrences of non-synonymous de-novo mutations in ORF1ab polyprotein subdivided by gene in 168 genomes obtained from 28 patients. NSP12 mutations were annotated relative to the reading frame post ribosomal frameshift, no mutations were observed within NSP12 prior to this loci (13,468). The first genome from each patient was considered to be day 0. The sampling periods and frequencies within the dataset was highly variable, 218 days was the longest time-period covered within the dataset but the majority were much shorter, the full details of the dataset are available in supplementary table 1. All mutations with cumulative frequencies ≥ 2 were labelled on-graph.

ORF1ab polyprotein genes, constituting the Non-structural proteins (NSPs) of SARS-CoV-2, demonstrated a larger number of recurrent mutations but still far fewer than in spike (Figure 4). Six DNMs were notable for their occurrence frequency; NSP3:T504P, NSP3:T820I, NSP3:P822L, NSP3:K977Q, NSP4:T295I, and NSP12:V792I. ORF1ab contained eighty-six out of the 195 DNMs observed, but only six of the total of twenty-one of the recurrent DNMs. ORF1ab constitutes more than two-thirds of the overall SARS-CoV-2 genome by length making the number of overall DNMs within the polyprotein disproportionately lower than would be expected if the distribution were random.
When DNMs observed within ORF1ab were clustered by AA loci the overall shape of the results remain broadly identical with two exceptions; NSP3:T504 and NSP3:P822 where their day 218 occurrences are raised to 3 and 4, respectively.
Figure 5: Spike mutational profiles of particular interest described by this study. Select spikes from late sequencing of 3 long-term Alpha infections shown as Spike schematics. Spike variants from WT Alpha, Delta and BA.1 Omicron shown for comparison. Mutations shown in grey are existing lineage-defining Alpha mutations. Mutations marked with an asterisk indicate mixed, but resolvable bases in the sequence.
<table>
<thead>
<tr>
<th>DNM Annotation</th>
<th>Frequency in DNM Occurrence Analysis</th>
<th>Frequency in COG-UK Dataset</th>
<th>Percentage of genome series in which DNM occurred</th>
<th>Percentage of Genomes in COG-UK with DNM</th>
</tr>
</thead>
<tbody>
<tr>
<td>S:E484K</td>
<td>8</td>
<td>3,437</td>
<td>28.57%</td>
<td>0.2180%</td>
</tr>
<tr>
<td>E:T30I</td>
<td>6</td>
<td>208</td>
<td>21.42%</td>
<td>0.0132%</td>
</tr>
<tr>
<td>M:H125Y</td>
<td>4</td>
<td>2,188</td>
<td>14.29%</td>
<td>0.1387%</td>
</tr>
<tr>
<td>S:Δ138 region</td>
<td>4</td>
<td>283,289</td>
<td>14.29%</td>
<td>17.9645%</td>
</tr>
<tr>
<td>NSP4:T295I</td>
<td>3</td>
<td>1,933</td>
<td>10.71%</td>
<td>0.1226%</td>
</tr>
<tr>
<td>S:Q493K</td>
<td>3</td>
<td>59</td>
<td>10.71%</td>
<td>0.0037%</td>
</tr>
<tr>
<td>S:Δ67 region</td>
<td>2</td>
<td>292,969</td>
<td>7.14%</td>
<td>18.5783%</td>
</tr>
<tr>
<td>S:S13I</td>
<td>2</td>
<td>211</td>
<td>7.14%</td>
<td>0.0134%</td>
</tr>
<tr>
<td>NSP12:V792I</td>
<td>2</td>
<td>10</td>
<td>7.14%</td>
<td>0.0006%</td>
</tr>
<tr>
<td>NSP3:P822L</td>
<td>2</td>
<td>28,410</td>
<td>7.14%</td>
<td>1.8016%</td>
</tr>
<tr>
<td>NSP3:T820I</td>
<td>2</td>
<td>442</td>
<td>7.14%</td>
<td>0.0280%</td>
</tr>
<tr>
<td>NSP3:T504P</td>
<td>2</td>
<td>18</td>
<td>7.14%</td>
<td>0.0011%</td>
</tr>
<tr>
<td>S:L452R</td>
<td>2</td>
<td>1,010,866</td>
<td>7.14%</td>
<td>64.1029%</td>
</tr>
<tr>
<td>S:Q498R</td>
<td>2</td>
<td>225</td>
<td>7.14%</td>
<td>0.0143%</td>
</tr>
<tr>
<td>S:E484G</td>
<td>2</td>
<td>46</td>
<td>7.14%</td>
<td>0.0029%</td>
</tr>
<tr>
<td>S:Δ243 region</td>
<td>2</td>
<td>546</td>
<td>7.14%</td>
<td>0.0346%</td>
</tr>
<tr>
<td>S:F486I</td>
<td>2</td>
<td>6</td>
<td>7.14%</td>
<td>0.0004%</td>
</tr>
<tr>
<td>S:G142V</td>
<td>2</td>
<td>1,361</td>
<td>7.14%</td>
<td>0.0863%</td>
</tr>
<tr>
<td>S:T95I</td>
<td>2</td>
<td>682,286</td>
<td>7.14%</td>
<td>43.2664%</td>
</tr>
<tr>
<td>NSP3:K977Q</td>
<td>2</td>
<td>391</td>
<td>7.14%</td>
<td>0.0248%</td>
</tr>
<tr>
<td>S:F490L</td>
<td>2</td>
<td>463</td>
<td>7.14%</td>
<td>0.0294%</td>
</tr>
</tbody>
</table>

Table 1: DNM Occurrence Frequencies for all Recurrent DNMs in this analysis and the COG-UK dataset (n=1,576,942). COG-UK dataset figures were generated using the dataset as it existed on 07/12/2021. Data was generated via CLIMB-Covid (Nicholls et al., 2021).

The relative frequencies for each recurrent mutation observed in the DNM occurrence analysis was compared to their prevalence within the COG-UK dataset (on the date 23/11/2021) (Table 1). As in the initial analysis S:E484K, E:T30I, and M:H125Y are noteworthy in their frequency especially compared to their low frequency in the larger COG-UK dataset.
<table>
<thead>
<tr>
<th>Mutation Annotation</th>
<th>Pango Lineage</th>
<th>UKHSA Label</th>
<th>WHO Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSP3:K977Q</td>
<td>P.1</td>
<td>VOC-21JAN-02</td>
<td>Gamma</td>
</tr>
<tr>
<td>NSP3:P822L</td>
<td>AV.1</td>
<td>VUI-21MAY-01</td>
<td>n/a</td>
</tr>
<tr>
<td>S:E484K</td>
<td>B.1.351</td>
<td>VOC-20DEC-02</td>
<td>Beta</td>
</tr>
<tr>
<td>S:E484K</td>
<td>B.1.525</td>
<td>VUI-21FEB-03</td>
<td>Eta</td>
</tr>
<tr>
<td>S:E484K</td>
<td>P.1</td>
<td>VOC-21JAN-02</td>
<td>Gamma</td>
</tr>
<tr>
<td>S:E484K</td>
<td>A.23.1</td>
<td>VUI-21FEB-01</td>
<td>n/a</td>
</tr>
<tr>
<td>S:E484K</td>
<td>AV.1</td>
<td>VUI-21MAY-01</td>
<td>n/a</td>
</tr>
<tr>
<td>S:E484K</td>
<td>B.1.1.318</td>
<td>VUI-21FEB-04</td>
<td>n/a</td>
</tr>
<tr>
<td>S:E484K</td>
<td>B.1.1.7 (with E484K)</td>
<td>VOC-21FEB-02</td>
<td>n/a</td>
</tr>
<tr>
<td>S:E484K</td>
<td>B.1.324.1</td>
<td>VUI-21MAR-01</td>
<td>n/a</td>
</tr>
<tr>
<td>S:E484K</td>
<td>P.3</td>
<td>VUI-21MAR-02</td>
<td>Theta</td>
</tr>
<tr>
<td>S:E484K</td>
<td>P.2</td>
<td>VUI-21JAN-01</td>
<td>Zeta</td>
</tr>
<tr>
<td>S:E484K</td>
<td>B.1.621</td>
<td>VUI-21JUL-01</td>
<td>n/a</td>
</tr>
<tr>
<td>S:L452R</td>
<td>B.1.617.2</td>
<td>VOC-21APR-02</td>
<td>Delta</td>
</tr>
<tr>
<td>S:L452R</td>
<td>B.1.617.1</td>
<td>VUI-21APR-01</td>
<td>Kappa</td>
</tr>
<tr>
<td>S:L452R</td>
<td>B.1.617.3</td>
<td>VUI-21APR-03</td>
<td>n/a</td>
</tr>
<tr>
<td>S:L452R</td>
<td>C.36.3</td>
<td>VUI-21MAY-02</td>
<td>n/a</td>
</tr>
<tr>
<td>S:Q498R</td>
<td>BA.1</td>
<td>VOC-21NOV-01</td>
<td>Omicron</td>
</tr>
<tr>
<td>S:T95I</td>
<td>AV.1</td>
<td>VUI-21MAY-01</td>
<td>n/a</td>
</tr>
<tr>
<td>S:T95I</td>
<td>B.1.1.318</td>
<td>VUI-21FEB-04</td>
<td>n/a</td>
</tr>
<tr>
<td>S:T95I</td>
<td>B.1.621</td>
<td>VUI-21JUL-01</td>
<td>n/a</td>
</tr>
<tr>
<td>S:Δ67 region/RDR1</td>
<td>B.1.1.7</td>
<td>VOC-20DEC-01</td>
<td>Alpha</td>
</tr>
<tr>
<td>S:Δ138 region/RDR2</td>
<td>B.1.1.7</td>
<td>VOC-20DEC-01</td>
<td>Alpha</td>
</tr>
<tr>
<td>S:Δ138 region/RDR2</td>
<td>AV.1</td>
<td>VUI-21MAY-01</td>
<td>n/a</td>
</tr>
<tr>
<td>S:Δ138 region/RDR2</td>
<td>B.1.1.318</td>
<td>VUI-21FEB-04</td>
<td>n/a</td>
</tr>
<tr>
<td>S:Δ243 region/RDR4</td>
<td>C.37</td>
<td>VUI-21JUN-01</td>
<td>Lambda</td>
</tr>
</tbody>
</table>

Table 2: Recurrent mutations which are variant defining based upon UKHSA variant definitions. Variant definitions were parsed from the UKHSA variant definition files available at: https://github.com/phe-genomics/variant_definitions. Lineages were called using pangolin (O’Toole et al., 2021b).
Each observed recurrent DNM was compared to the UKHSA VOC/VUI definition files (Table 2). S:E484K was the most frequent DNM to appear in VOC/VUI definitions with eleven appearances, then S:L452R with four, then S:T95I and S:Δ138/RDR2 region with three each, followed by NSP3:K977Q, NSP3:P822L, S:Q498R, S:Δ67/RDR1 region, and S:Δ243/RDR4 region with one each. Of the twenty-one recurrent DNMs observed in the analysis nine of them are considered defining mutations for a VOC/VUI.
Discussion

Not all mutations are discussed in detail, while a literature search has been performed for every recurrent DNM only those with sufficient literature available for discussion to be informative were included below.

S-Gene - Receptor Binding Domain Recurrent Mutations

The frequency of RBD DNMs observed in this analysis is a significant finding; the RBD is a relatively small region of the SARS-CoV-2 genome making up less than two percent of the genome by length, but these account for 17 percent of all DNMs observed (Figure 1). It is clear that RBD mutations were the most strongly selected for in the immunocompromised patients included within the dataset.

The sharp rise of S:E484K occurrences early in the period is biased due to the data from (Jensen et al., 2021) as a result of their sampling strategy and research focus. (Jensen et al., 2021) specifically discussed the emergence of S:E484K in long-term immunocompromised patients and published short periods of surveillance of these cases when the patients in question had significantly longer shedding periods to demonstrate this. However, even if this study is excluded S:E484K remains the most frequently occurring DNM within spike.

The high frequency of the S:E484K occurrences is suggestive of a strong selective pressure; this is further demonstrated by the total of twelve DNMs observed at the S:484 loci. The two occurrences of S:E484G in the dataset also suggest that the glycine substitution is subject to differing selection pressures than the lysine substitution in S:E484K although this may be host dependent. In one of the two occurrences of S:E484G this change was transient and was replaced by S:E484K. There are two possible explanations for this observation; a secondary mutation, or both mutations occurred within the patient and the S:E484K subpopulation outcompeted the S:E484G population to become dominant. There is no single nucleotide change by which a G -> K AA change might occur, supporting the second possibility. If the second explanation is correct it would suggest that S:484 mutations are selected for generally. The large difference between the frequency of S:E484K in this dataset compared to the national COG-UK dataset further suggests that the selection pressures which caused S:E484K to be so frequent within this analysis are not true of the majority of hosts (Table 1). S:E484K is also considered a defining mutation for a large number of variants, further indicating a
A strong selective pressure for S:E484K was also observed by Zahradník et al., (2021) who discovered using an *in vitro* experimental evolution model, that >70% of clones in one library gained S:E484K and S:N501Y which were associated with a significant increase in ACE2 affinity. Furthermore they observed the occurrence of the mutation S:Q498R alongside S:N501Y in two repeats, this combination was observed to lead to significantly greater affinity to ACE2 compared to both wild-type and Alpha which rose further alongside S:E484K. This combination was only observed within a single patient (pt 19) although the combination E484G, Q498R and N501Y did arise in a further patient (pt 17), in both cases the infections were Alpha and therefore already possessed S:N501Y. At the time Zahradník et al was published this constellation of mutations had not been observed in wild virus but with the emergence of Omicron this combination has become significantly more frequent (albeit with E484A rather than E484K).

The low occurrence frequency of S:N501Y compared to that observed by (Zahradník et al., 2021) is also notable but is partly explained by its high (nine out of twenty-eight) day zero frequency in the genome series’, due to the high amount of long-term Alpha infections included in this study. When DNMs were clustered by AA loci S:501 was highlighted as recurrent however.

Another notable observation is the two *de novo* occurrences of S:L452R (a defining mutation of Delta, Kappa and Epsilon variants) which aids both immune evasion and ACE2 affinity (Motozono et al., 2021).

S:Q493K has previously been identified by Huang et al, (2021) as a highly beneficial adaptation to a mouse host, improving spike binding affinity to murine ACE2 (Huang et al., 2021), its rarity in the overall SARS-CoV-2 population (58 in COG-UK dataset) suggests that it is not strongly selected for in a human host generally. The three occurrences in this dataset may suggest that S:Q493K does confer a benefit to the virus within the context of a long-term infection but not in transient infection. A highly similar mutation, S:Q493R, is a defining mutation of the Omicron variant.
S:F486I has been observed to decrease the affinity of some neutralising antibodies to spike protein (Xu et al., 2021), and may decrease the affinity of spike to ACE2 (Clark et al., 2021). S:F486I has furthermore been associated with mink adaptation (Zhou et al., 2021). S:490L has been observed to reduce the affinity of multiple mAbs as well as decreasing the neutralisation sensitivity of pseudovirus to convalescent sera, however it does not appear to have an impact on viral infectivity (Li et al., 2020). It is noteworthy that a large number of mutations described in this present study are associated with enhanced human ACE2 affinity including Q493K, Q498R and N501Y (Starr et al., 2020).

When AA loci clustering was performed recurrent DNMs at S:330 and S:440 were observed.

Finally, although most of this study has considered mutations in isolation, several of the late stage long-term infections showed interesting combinations of mutations, particularly within Spike (figure 5). Patient 19 for example was an Alpha infection that had picked up a large number of mutations, many of which were in common with, or similar to Omicron, for example A67D, G142V, T95I, Δ210/L212I, E484K, and Q498R. A further case, patient 17 also contained E484G and Q498R alongside the Alpha lineage defining mutation, N501Y and pt 27 contained T95I, a further deletion at S:Δ138 region and G496S, in common with Omicron.

**S-Gene N-terminal Domain Recurrent Mutations**

T95I has been shown to bind of the human Tyrosine-protein kinase receptor UFO (AXL) and it has been suggested by (Singh et al., 2021) that AXL facilitates SARS-CoV-2 cell entry to the same extent as ACE2 in AXL overexpressed cell culture. NTD also has a substantial role in the antigenicity of spike with multiple escape mutations identified in this domain (Harvey et al., 2021).

All recurrent deletions within the SARS-CoV-2 genome were observed within the NTD (S:Δ67 region/RDR1, S:Δ138/RDR2 region, and S:Δ243/RDR4 region). Deletions within the S:69-70 region are commonly observed (McCarthy et al., 2021; Meng et al., 2021). Meng et al, (2021) characterised the common S:Δ69-70 deletion as contributing to infectivity by improving incorporation of cleaved spike protein into virions and possibly has a compensatory effect on mutations in the RBD associated with Ab escape such as N439K and Y453F. Of the two observations of deletions within the S:67-70 region one was S:Δ69-70 whereas the other was S:Δ67 which
has not been commonly observed, but it is notable that the genome series in which S:Δ67 was observed already possessed S:Δ69 at day 0. S:Δ69-70 is also a defining mutation of the Alpha and Omicron variants and is responsible for the S-gene target failure observed in the PCR testing of alpha variant samples with TaqPath SARS-CoV-2 PCR kits (Kidd et al., 2021).

De novo occurrences of slightly differing deletions within the S:Δ138/RDR2 region were observed 4 times. This region makes up part of the “NTD antigenic supersite” which the majority of neutralising antibodies against the NTD target (McCallum et al., 2021b). S:Δ140 has consequently been associated with a significant decrease in Ab neutralisation (Andreano et al., 2021; Liu et al., 2021). Based upon the high number of occurrences, it appears likely that deletions in this region confer some benefit to the virus during long-term infections. As with S:N501Y, as well as S:Δ67, it is worth noting a substantial proportion of long-term infections already carried deletions in the S:Δ138 region at day 0 due to being the Alpha variant.

Two occurrences of S:Δ243, another NTD supersite mutation, were also observed, another deletion which has been demonstrated to decrease Ab neutralisation in vitro (McCallum et al., 2021b; McCarthy et al., 2021).

**S-Gene Signal Peptide Recurrent Mutations**

The single recurrent signal peptide DMN, S:S13I, has been previously shown to mediate a shift of the cleavage site of the signal peptide which in turn facilitates immune evasion by causing a significant re-arrangement of the NTD antigenic supersite and its constituent internal disulphide bonding (McCallum et al., 2021a, 2021b).

**E-Gene Recurrent Mutations**

The most frequent de novo mutation observed outside of the spike gene is Envelope:T30I (the second most frequent mutation overall after S:E484X). This mutation was observed by Chaudhry et al, (2020) in a cell-culture passage experiment, where it conferred a growth advantage in Calu-3 cells but slowed growth in Vero E6 cells (Chaudhry et al., 2020).

The high frequency of E:T30I is strongly suggestive of a selective pressure during long-term infections and further suggests that the conditions experienced by the virus in immunocompromised patients may exist in a
similar selective environment as cell culture, potentially due to a lack of stability needed for transmission. The significant enrichment of E:T30I in this analysis compared to the COG-UK dataset (Table 2) suggests that E:T30I may be a deleterious mutation within the circulating SARS-CoV-2 population. A single variant lineage, B.1.616, does contain E:T30I as a lineage-defining mutation. Interestingly, B.1.616 was associated with an extremely localised, largely nosocomial-associated outbreak, suggesting the possibility this may have been the emergence of a virus from a long-term infection (Fillâatre et al., 2021). This also raises the hypothetical possibility that E:T30I may be considered a marker of long-term SARS-CoV-2 infections. Further study is necessary to determine the phenotypic effect of this mutation and its role in influencing within and between host fitness.

**ORF1ab-NSP3 Recurrent Mutations**

Literature concerning mutations in ORF1ab is generally observational rather than experimental due to the current lack of tractable models to study them *in vitro*. The concentration of higher frequency mutations within the NSP3 gene is not surprising considering it is the largest gene within the ORF1ab polyprotein and is known to be a bulky, modular protein which may have some flexible linker regions which are fairly hypermutatable. Stanevich et al. identified NSP3:T504P as a mutation associated with cytotoxic T cell epitope immune escape (Stanevich et al., 2021).

**Conclusions**

This work sought to determine recurrent mutations across the SARS-CoV-2 genome associated with long-term infections in immunodeficient patients. This study has several notable limitations: importantly a significant publication bias is likely to be present which may overemphasis the importance of some mutations. S:E484K especially is affected by this, the six genome series obtained from Jensen et al., (2021) were published to demonstrate the emergence of S:E484K within immunocompromised patients. Further work will attempt to avoid this by utilising less biased sampling strategies from long-term infected patients, requiring a prospective study design that aims to regularly sample genomes from long-term infected patients.
The majority of recurrently observed DNMs have been associated with immune escape, increased ACE2 affinity, or improved viral packaging and are generally not highly prevalent within the wider SARS-CoV-2 population (with the exception of some SARS-CoV-2 variants).

These factors suggest that the conditions during long-term infections at least partly select for mutations which aid the virus with intra-host replication and persistence as opposed to the general SARS-CoV-2 population, where mutations which aid inter-host transmission are more strongly selected for. E:T30I in particular is worthy of further study as a potential marker of long-term SARS-CoV-2 infections.

However, the large number of occurrences overlapping with variant defining mutations observed does indicate that patients within this category should not be discounted as a potential source of previous, or indeed future variants. The potential of mutations which aid cell-cell transmission within the host or improve viral packaging may affect virulence and any mutations within this category which do not impact viral transmissibility could have a significant impact. This is highly relevant as many of the most abundant mutations described in this dataset are found across many variant lineages. Furthermore, its possible sub-neutralising levels of antibodies which may be present in some cases (either homologous or from heterologous convalescent or monoclonal antibody treatments) could be selecting for the acquisition of antigenic mutations seen in several of these cases (Kemp et al., 2021).

At present it is unresolved where SARS-CoV-2 variants emerge from. One prevailing hypothesis is that some variants emerged from long-term chronic infections, generating novel advantageous combinations of mutations without the stringent selection pressure of transmission, eventually resulting in an outbreak and onward transmission. We have compared common mutations arising during chronic infections and described how many are shared with SARS-CoV-2 variant lineages. Furthermore we present evidence, based on a rare mutational signature, that the French B.1.616 variant lineage arose from a direct and recent spillover from a chronic infection. Overall the data presented here is consistent and supportive of the chronic infection hypothesis of SARS-CoV-2 variant emergence. Therefore we suggest identifying and curing chronic infections, preferably with combined antiviral therapy as would be used for more traditionally chronic viruses (HIV, HCV), both to the infected individual, but also to global health. Intrahost variation of SARS-CoV-2 is likely to play a
significant role within this patient group however the lack of raw data availability for the majority of the samples within this dataset makes this challenging (Chaudhry et al., 2020).

We anticipate this dataset will be maintained as a public resource to enable the study of long-term SARS-CoV-2 infections in immunodeficient patients for as long as it is deemed relevant to enable other researchers to contribute to this understudied, highly important, patient group (https://github.com/BioWilko/recurrent-sars-cov-2-mutations/blob/main/dataset/mutation_calls.csv).
References


Acknowledgments

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) [grant code: MC_PC_19027], and Genome Research Limited, operating as the Wellcome Sanger Institute.

The COG-UK study protocol was approved by the Public Health England Research Ethics Governance Group (reference: R&D NR0195). Authors only had access to anonymised data. No individual patient consent was required.

Queen Elizabeth Hospital, University Hospitals Birmingham, Birmingham B15 2TH, UK.

- Mark Garvey, Anna Casey, Liz Ratcliffe, Husam Osman
- Contact: Anna.Casey@uhb.nhs.uk

- Contact: Manuela Cernadas


- Oksana Stanevich, Evgenia Alekseeva, Maria Sergeeva, Artem Fadeev, Kseniya Komissarova, Anna Ivanova, Tamara Simakova, Kirill Vasilyev, Anna-Polina Shurygina, Marina Stukova, Ksenia Safina, Elena Nabieva, Sofya Garushyants, Galya Klink, Evgeny Bakin, Julia Zabutova, Anastasia Kholodnaia, Olga Lukina, Irina Skorokhod, Viktoria Ryabchikova, Nadezhda Medvedeva, Dmitry Lioznov, Daria Danilenko, Dmitriy Chudakov, Andrey Komissarov, Georgii Bazykin.

- Contact: evg.alekseeva93@gmail.com


- Elham Khatamzas, Alexandra Rehn, Maximilian Muenchhoff, Johannes Hellmuth, Erik Gaitzsch, Tobias Weiglein, Enrico Georgi, Clemens Scherer, Stephanie Stecher, Oliver Weigert, Philipp Girl, Sabine Zange, Oliver T. Keppler, Joachim Stemmler, Michael von Bergwelt-Baildon, Roman Wölfel, Markus Antwerpen.

- Contact: elham.khatamzas@med.uni-muenchen.de

- Vítor Borges, Joana Isidro, Mário Cunha, Daniela Cochicho, Luis Martins, Luis Banha, Margarida Figueiredo, Leonor Rebelo, Maria Céu Trindade, Silvia Duarte, Luís Vieira, Maria João Alves, Inês Costa, Raquel Guiomar, Madalena Santos, Rita Corté-Real, André Dias, Diana Póvoas, João Cabo, Carlos Figueiredo, Maria José Manata, Fernando Mel tez, Maria Gomes da Silva, João Paulo Gomes.

- Contact: j.paulo.gomes@insa.min-saude.pt

Virology Department, NHS East and South East London Pathology Partnership, Royal London Hospital, Barts Health NHS Trust:

- Beatrix Kele, Kathryn Harris, Theresa Cutino-Moguel, Dola Owoyemi, Shahiba Sultanam, Abril Romero.

- Contact: beatrix.kele@nhs.net


- Contact: cflores@ull.edu.es


- Contact: bjoern-erikole.jensen@med.uni-duesseldorf.de


- Sebastian Weigang, Jonas Fuchs, Gert Zimmer, Daniel Schnepf, Lisa Kern, Julius Beer, Hendrik Luxenburger, Jakob Ankerhold, Valeria Falcone, Janine Kemming, Maike Hofmann, Robert Thimme,
Christoph Neumann-Haefelin, Svenja Ulferts, Robert Grosse, Daniel Hornuss, Yakup Tanriver, Siegbert Rieg, Dirk Wagner, Daniela Huzly, Martin Schwemmle, Marcus Panning, Georg Kochs.

- **Contact:** marcus.panning@uniklinik-freiburg.de & georg.kochs@uniklinik-freiburg.de
The COVID-19 Genomics UK (COG-UK) consortium

June 2021 V.1

Funding acquisition, Leadership and supervision, Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, Software and analysis tools, and Visualisation:

Samuel C Robson 13, 84

Funding acquisition, Leadership and supervision, Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, and Software and analysis tools:

Thomas R Connor 11, 74 and Nicholas J Loman 43

Leadership and supervision, Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, Software and analysis tools, and Visualisation:

Tanya Golubchik 5

Funding acquisition, Leadership and supervision, Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, and Software and analysis tools:

Rocio T Martinez Nunez 46

Funding acquisition, Leadership and supervision, Project administration, Samples and logistics, Sequencing and analysis, and Software and analysis tools:

David Bonsall 5

Funding acquisition, Leadership and supervision, Project administration, Sequencing and analysis, Software and analysis tools, and Visualisation:

Andrew Rambaut 104

Funding acquisition, Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, and Software and analysis tools:

Luke B Snell 12

Leadership and supervision, Metadata curation, Project administration, Samples and logistics, Software and analysis tools, and Visualisation:
Rich Livett

Funding acquisition, Leadership and supervision, Metadata curation, Project administration, and Samples and logistics:

Catherine Ludden

Funding acquisition, Leadership and supervision, Metadata curation, Samples and logistics, and Sequencing and analysis:

Sally Corden and Eleni Nastouli

Funding acquisition, Leadership and supervision, Metadata curation, Sequencing and analysis, and Software and analysis tools:

Gaia Nebbia

Funding acquisition, Leadership and supervision, Project administration, Samples and logistics, and Sequencing and analysis:

Ian Johnston

Leadership and supervision, Metadata curation, Project administration, Samples and logistics, and Sequencing and analysis:

Katrina Lythgoe, M. Estee Torok and Ian G Goodfellow

Leadership and supervision, Metadata curation, Project administration, Samples and logistics, and Visualisation:

Jacqui A Prieto and Kordo Saeed

Leadership and supervision, Metadata curation, Project administration, Sequencing and analysis, and Software and analysis tools:

David K Jackson

Leadership and supervision, Metadata curation, Samples and logistics, Sequencing and analysis, and Visualisation:

Catherine Houlihan

Leadership and supervision, Metadata curation, Sequencing and analysis, Software and analysis tools, and Visualisation:

Dan Frampton
Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, and Software and analysis tools:

William L Hamilton 19 and Adam A Witney 41

Funding acquisition, Samples and logistics, Sequencing and analysis, and Visualisation:

Giselda Bucca 101

Funding acquisition, Leadership and supervision, Metadata curation, and Project administration:

Cassie F Pope 40, 41

Funding acquisition, Leadership and supervision, Metadata curation, and Samples and logistics:

Catherine Moore 74

Funding acquisition, Leadership and supervision, Metadata curation, and Sequencing and analysis:

Emma C Thomson 53

Funding acquisition, Leadership and supervision, Project administration, and Samples and logistics:

Ewan M Harrison 116, 102

Funding acquisition, Leadership and supervision, Sequencing and analysis, and Visualisation:

Colin P Smith 101

Leadership and supervision, Metadata curation, Project administration, and Sequencing and analysis:

Fiona Rogan 77

Leadership and supervision, Metadata curation, Project administration, and Samples and logistics:

Shaun M Beckwith 6, Abigail Murray 6, Dawn Singleton 6, Kirstine Eastick 37, Liz A Sheridan 98, Paul Randell 99, Leigh M Jackson 105, Cristina V Ariani 116 and Sónia Gonçalves 116

Leadership and supervision, Metadata curation, Samples and logistics, and Sequencing and analysis:

Derek J Fairley 3, 77, Matthew W Loose 18 and Joanne Watkins 74
Leadership and supervision, Metadata curation, Samples and logistics, and Visualisation:
Samuel Moses

Leadership and supervision, Metadata curation, Sequencing and analysis, and Software and analysis tools:
Sam Nicholls, Matthew Bull and Roberto Amato

Leadership and supervision, Project administration, Samples and logistics, and Sequencing and analysis:
Darren L Smith

Leadership and supervision, Sequencing and analysis, Software and analysis tools, and Visualisation:
David M Aanensen and Jeffrey C Barrett

Metadata curation, Project administration, Samples and logistics, and Sequencing and analysis:
Dinesh Aggarwal, James G Shepherd, Martin D Curran and Surendra Parmar

Metadata curation, Project administration, Sequencing and analysis, and Software and analysis tools:
Matthew D Parker

Metadata curation, Samples and logistics, Sequencing and analysis, and Software and analysis tools:
Catryn Williams

Metadata curation, Samples and logistics, Sequencing and analysis, and Visualisation:
Sharon Glaysher

Metadata curation, Sequencing and analysis, Software and analysis tools, and Visualisation:
Anthony P Underwood, Matthew Bashton, Nicole Pacchiarini, Katie F Loveson and Matthew Byott

Project administration, Sequencing and analysis, Software and analysis tools, and Visualisation:
Alessandro M Carabelli
Funding acquisition, Leadership and supervision, and Metadata curation:
Kate E Templeton 56, 104

Funding acquisition, Leadership and supervision, and Project administration:
Thushan I de Silva 109, Dennis Wang 109, Cordelia F Langford 116 and John Sillitoe 116

Funding acquisition, Leadership and supervision, and Samples and logistics:
Rory N Gunson 55

Funding acquisition, Leadership and supervision, and Sequencing and analysis:
Simon Cottrell 74, Justin O’Grady 75, 103 and Dominic Kwiatkowski 116, 108

Leadership and supervision, Metadata curation, and Project administration:
Patrick J Lillie 37

Leadership and supervision, Metadata curation, and Samples and logistics:
Nicholas Cortes 33, Nathan Moore 33, Claire Thomas 33, Phillipa J Burns 37, Tabitha W Mahungu 80 and Steven Liggett 86

Leadership and supervision, Metadata curation, and Sequencing and analysis:
Angela H Beckett 13, 81 and Matthew TG Holden 73

Leadership and supervision, Project administration, and Samples and logistics:
Lisa J Levett 34, Husam Osman 70, 35 and Mohammed O Hassan-Ibrahim 99

Leadership and supervision, Project administration, and Sequencing and analysis:
David A Simpson 77

Leadership and supervision, Samples and logistics, and Sequencing and analysis:
Meera Chand 72, Ravi K Gupta 102, Alistair C Darby 107 and Steve Paterson 107

Leadership and supervision, Sequencing and analysis, and Software and analysis tools:
Leadership and supervision, Sequencing and analysis, and Visualisation:
Louise Aigrain and Andrew R Bassett

Metadata curation, Project administration, and Samples and logistics:
Nick Wong, Yusri Taha, Michelle J Erkiert and Michael H Spencer Chapman

Metadata curation, Project administration, and Sequencing and analysis:
Rebecca Dewar and Martin P McHugh

Metadata curation, Project administration, and Software and analysis tools:
Siddharth Mookerjee

Metadata curation, Project administration, and Visualisation:
Stephen Aplin, Matthew Harvey, Thea Sass, Helen Umpleby and Helen Wheeler

Metadata curation, Samples and logistics, and Sequencing and analysis:
James P McKenna, Ben Warne, Joshua F Taylor, Yasmin Chaudhry, Rhys Izuagbe, Aminu S Jahun, Gregory R Young, Claire McMurray, Clare M McCann, Andrew Nelson and Scott Elliott

Metadata curation, Samples and logistics, and Visualisation:
Hannah Lowe

Metadata curation, Sequencing and analysis, and Software and analysis tools:
Anna Price, Matthew R Crown, Sara Rey, Sunando Roy and Ben Temperton

Metadata curation, Sequencing and analysis, and Visualisation:
Sharif Shaaban and Andrew R Hesketh

Project administration, Samples and logistics, and Sequencing and analysis:
Kenneth G Laing, Irene M Monahan and Judith Heaney

Project administration, Samples and logistics, and Visualisation:
Emanuela Pelosi, Siona Silviera and Eleri Wilson-Davies

Samples and logistics, Software and analysis tools, and Visualisation:
Helen Fryer

Sequencing and analysis, Software and analysis tools, and Visualization:
Helen Adams, Louis du Plessis, Rob Johnson, William T Harvey, Joseph Hughes, Richard J Orton, Lewis G Spurgin, Yann Bourgeois, Chris Ruis, Áine O’Toole, Marina Gourtovaia and Theo Sanderson

Funding acquisition, and Leadership and supervision:
Christophe Fraser, Jonathan Edgeworth, Judith Breuer, Stephen L Michell and John A Todd

Funding acquisition, and Project administration:
Michaela John and David Buck

Leadership and supervision, and Metadata curation:
Kavitha Gajee and Gemma L Kay

Leadership and supervision, and Project administration:
Sharon J Peacock and David Heyburn

Leadership and supervision, and Samples and logistics:
Katie Kitchman, Alan McNally, David T Pritchard, Samir Dervisevic, Peter Muir, Esther Robinson, Barry B Vipond, Newara A Ramadan, Christopher Jeanes, Danni Weldon, Jana Catalan and Neil Jones

Leadership and supervision, and Sequencing and analysis:
Ana da Silva Filipe, Chris Williams, Marc Fuchs, Julia Miskelly, Aaron R Jeffries, Karen Oliver and Naomi R Park
Metadata curation, and Samples and logistics:

Amy Ash 1, Cherian Koshy 1, Magdalena Barrow 7, Sarah L Buchan 7, Anna Mantzouratou 7, Gemma Clark 15, Christopher W Holmes 16, Sharon Campbell 17, Thomas Davis 21, Ngee Keong Tan 22, Julianne R Brown 29, Kathryn A Harris 29, 2, Stephen P Kidd 33, Paul R Grant 34, Li Xu-McCrae 35, Alison Cox 38, 63, Pinglawathee Madona 38, 63, Marcus Pond 38, 63, Paul A Randell 38, 63, Karen T Withell 48, Cheryl Williams 51, Clive Graham 60, Rebecca Denton-Smith 62, Emma Swindells 62, Robyn Turnbull 62, Tim J Sloan 67, Andrew Bosworth 70, 35, Stephanie Hutchings 70, Hannah M Pymont 70, Anna Casey 76, Liz Ratcliffe 76, Christopher R Jones 79, 105, Bridget A Knight 79, 105, Tanzina Haque 80, Jennifer Hart 80, Dianne Irish-Tavares 80, Eric Witele 80, Craig Mower 86, Louisa K Watson 86, Jennifer Collins 89, Gary Eltringham 89, Dorian Crudgington 98, Ben Macklin 98, Miren Iturriza-Gomara 107, Anita O Lucaci 107 and Patrick C McClure 113

Metadata curation, and Sequencing and analysis:

Matthew Carlile 18, Nadine Holmes 18, Christopher Moore 18, Nathaniel Storey 79, Stefan Rooke 73, Gonzalez Yebra 73, Noel Craine 74, Malorie Perry 74, Nabil-Fareed Alikhan 75, Stephen Bridgett 77, Kate F Cook 84, Christopher Fearn 84, Salman Goudarzi 84, Ronan A Lyons 88, Thomas Williams 104, Sam T Haldenby 107, Jillian Durham 116 and Steven Leonard 116

Metadata curation, and Software and analysis tools:

Robert M Davies 116

Project administration, and Samples and logistics:


Project administration, and Sequencing and analysis:

Joshua Quick 43

Project administration, and Software and analysis tools:

Radoslaw Poplawski 43

Samples and logistics, and Sequencing and analysis:


Samples and logistics, and Software and analysis tools:
Igor Starinskij 53

Sequencing and analysis, and Software and analysis tools:

Sequencing and analysis, and Visualisation:
MacGregor Cox 20

Software and analysis tools, and Visualisation:

Leadership and supervision:
Tanya Curran 3, Sian Morgan 10, Patrick Maxwell 20, Ken Smith 20, Sahar Eldirdiri 21, Anita Kenyon 21, Alison H Holmes 38, 57, James R Price 38, 57, Tim Wyatt 69, Alison E Mather 75, Timofey Skvortsov 77 and John A Hartley 96

Metadata curation:
Martyn Guest 11, Christine Kitchen 11, Ian Merrick 11, Robert Munn 11, Beatrice Bertolusso 33, Jessica Lynch 33, Gabrielle Vernet 33, Stuart Kirk 34, Elizabeth Wastnedge 56, Rachael Stanley 58, Giles Idle 64, Declan T Bradley 69, 77, Jennifer Poyner 79 and Matilde Mori 110

Project administration:
Samples and logistics:


Sequencing and analysis:


Software and analysis tools:

Will Rowe 43, Amy Gaskin 74, Thanh Le-Viet 75, James Bonfield 116, Jenniffer Liddle 116 and Andrew Whitwham 116
1 Barking, Havering and Redbridge University Hospitals NHS Trust, 2 Barts Health NHS Trust, 3 Belfast Health & Social Care Trust, 4 Betsi Cadwaladr University Health Board, 5 Big Data Institute, Nuffield Department of Medicine, University of Oxford, 6 Blackpool Teaching Hospitals NHS Foundation Trust, 7 Bournemouth University, 8 Cambridge Stem Cell Institute, University of Cambridge, 9 Cambridge University Hospitals NHS Foundation Trust, 10 Cardiff and Vale University Health Board, 11 Cardiff University, 12 Centre for Clinical Infection and Diagnostics Research, Department of Infectious Diseases, Guy's and St Thomas' NHS Foundation Trust, 13 Centre for Enzyme Innovation, University of Portsmouth, 14 Centre for Genomic Pathogen Surveillance, University of Oxford, 15 Clinical Microbiology Department, Queens Medical Centre, Nottingham University Hospitals NHS Trust, 16 Clinical Microbiology, University Hospitals of Leicester NHS Trust, 17 County Durham and Darlington NHS Foundation Trust, 18 Deep Seq, School of Life Sciences, Queens Medical Centre, University of Nottingham, 19 Department of Infectious Diseases and Microbiology, Cambridge University Hospitals NHS Foundation Trust, 20 Department of Medicine, University of Cambridge, 21 Department of Microbiology, Kettering General Hospital, 22 Department of Microbiology, South West London Pathology, 23 Department of Zoology, University of Oxford, 24 Division of Virology, Department of Pathology, University of Cambridge, 25 East Kent Hospitals University NHS Foundation Trust, 26 East Suffolk and North Essex NHS Foundation Trust, 27 East Sussex Healthcare NHS Trust, 28 Gateshead Health NHS Foundation Trust, 29 Great Ormond Street Hospital for Children NHS Foundation Trust, 30 Great Ormond Street Institute of Child Health (GOS ICH), University College London (UCL), 31 Guy's and St. Thomas' Biomedical Research Centre, 32 Guy's and St. Thomas' NHS Foundation Trust, 33 Hampshire Hospitals NHS Foundation Trust, 34 Health Services Laboratories, 35 Heartlands Hospital, Birmingham, 36 Hub for Biotechnology in the Built Environment, Northumbria University, 37 Hull University Teaching Hospitals NHS Trust, 38 Imperial College Healthcare NHS Trust, 39 Imperial College London, 40 Infection Care Group, St George's University Hospitals NHS Foundation Trust, 41 Institute for Infection and Immunity, St George's University of London, 42 Institute of Biodiversity, Animal Health & Comparative Medicine, 43 Institute of Microbiology and Infection, University of Birmingham, 44 Isle of Wight NHS Trust, 45 King's College Hospital NHS Foundation Trust, 46 King's College London, 47 Liverpool Clinical Laboratories, 48 Maidstone and Tunbridge Wells NHS Trust, 49 Manchester University NHS Foundation Trust, 50 Microbiology Department, Buckinghamshire Healthcare NHS Trust, 51 Microbiology, Royal Oldham Hospital, 52 MRC Biostatistics Unit, University of Cambridge, 53 MRC-University of Glasgow Centre for Virus Research, 54 Newcastle University, 55 NHS Greater Glasgow and Clyde, 56 NHS Lothian, 57 NIHR Health Protection Research Unit in HCAI and AMR, Imperial College London, 58 Norfolk and Norwich University Hospitals NHS Foundation Trust, 59 Norfolk County Council, 60 North Cumbria Integrated Care NHS Foundation Trust, 61 North Middlesex University Hospital NHS Trust, 62 North Tees and Hartlepool NHS Foundation Trust, 63 North West London Pathology, 64 Northumbria Healthcare NHS Foundation Trust, 65 Northumbria University, 66 NU-OMICS, Northumbria University, 67 Path Links, Northern Lincolnshire and Goole NHS Foundation Trust, 68 Portsmouth Hospitals University NHS Trust, 69 Public Health Agency, Northern Ireland, 70 Public Health England, 71 Public Health England, Cambridge, 72 Public Health England, Colindale, 73 Public Health Scotland, 74 Public Health Wales, 75 Quadram Institute Bioscience, 76 Queen Elizabeth Hospital, Birmingham, 77 Queen's University Belfast, 78 Royal Brompton and Harefield Hospitals, 79 Royal Devon and Exeter NHS Foundation Trust, 80 Royal Free London NHS Foundation Trust, 81 School of Biological Sciences, University of Portsmouth, 82 School of Health Sciences, University of Southampton, 83 School of Medicine, University of Southampton, 84 School of Pharmacy & Biomedical Sciences, University of Portsmouth, 85 Sheffield Teaching Hospitals NHS Foundation Trust, 86 South Tees Hospitals NHS Foundation Trust, 87 Southwest Pathology Services, 88 Swansea University, 89 The Newcastle upon Tyne Hospitals NHS Foundation Trust, 90 The Queen Elizabeth Hospital King's Lynn NHS Foundation Trust, 91 The Royal Marsden NHS Foundation Trust, 92 The Royal Wolverhampton NHS Trust, 93 Turnkey Laboratory, University of Birmingham, 94 University College London Division of Infection and Immunity, 95 University College London Hospital Advanced Pathogen Diagnostics Unit, 96 University College London Hospitals NHS Foundation Trust, 97 University Hospital Southampton NHS Foundation Trust, 98 University Hospitals Dorset NHS Foundation Trust, 99 University Hospitals Sussex NHS Foundation Trust, 100 University of Birmingham, 101 University of Brighton, 102 University of Cambridge, 103 University of East Anglia, 104 University of Edinburgh, 105 University of Exeter, 106 University of Kent, 107 University of Liverpool, 108 University of Oxford, 109 University of Sheffield, 110 University of Southampton, 111 University of St Andrews, 112 Viapath, Guy's and St Thomas' NHS Foundation Trust, and King's College Hospital NHS Foundation Trust, 113 Virology, School of Life Sciences, Queens Medical Centre, University of Nottingham, 114 Watford General Hospital, 115 Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, 116 Wellcome Sanger Institute, 117 West of Scotland Specialist Virology Centre, NHS Greater Glasgow and Clyde, 118 Whittington Health NHS Trust