COVID-19 Genomics UK (COG-UK) Consortium

Report #8 - 11th June 2020

This report is provided at the request of SAGE and includes information on the ongoing state of the investigations being carried out. It should not be considered formal or informal advice. The conclusions of the ongoing scientific studies may be subject to change as further evidence becomes available and as such any firm conclusions would be premature.

Executive Summary

- The scale of the data generated by COG-UK continues to increase weekly, with more that 25K SARS-CoV-2 genomes now sequenced, corresponding to ~55% of the global total.
- A preliminary analysis of SARS-CoV-2 importation and establishment in the UK has identified at least 1356 independently-introduced UK transmission lineages and revealed that the rate and source of introductions changed substantially over time. Most introductions likely occurred during a window in which high inbound travel volumes coincided with increasing case numbers in several European countries.
- An analysis of introductions into Scotland estimated at least 113 introductions of SARS-CoV-2 during March 2020, revealed that undetected introductions occurred prior to the first known case, and identified an apparent shift from travel-associated introduction to sustained community transmission after 11 days.
- An updated survey of six London care homes provides evidence for multiple introductions and subsequent spread of SARS-CoV-2 within care homes, but could not provide conclusive evidence that cases of a lineage present in multiple sites were linked and caused by spread between care homes.

COG-UK update

Across the 17 active sequencing sites, the total number of SARS-CoV-2 genomes sequenced by COG-UK now stands at 25,052 (Figure 1a), constituting ~55% of the global total number of SARS-CoV-2 genomes (Figure 2b) and continuing to grow by ~10% every week

The scale of the COG-UK dataset is unparalleled and has enabled analyses that overcome the constraints imposed by low genomic diversity in the smaller sample pool available to most other countries (See "Preliminary analysis of SARS-CoV-2 importation & establishment in the UK", below).

To ensure that clinicians, public health workers and researchers have access to the most recent data when analysing their samples, work is underway to move the relevant COG-UK bioinformatic pipelines towards a 24-hour data update (rather than the present weekly one). Going forward, the ability for users to incorporate their samples directly into existing phylogenetic trees is being developed and will increase the ease with which samples can be analysed in the context of the wider UK dataset.

Work is underway to design a strategy for handling plates received from the Lighthouse national testing centres. These plates have a relatively few positive samples present and are currently being stored at the Wellcome Sanger Institute, but could be sequenced using spare capacity at regional COG-UK sequencing centres.

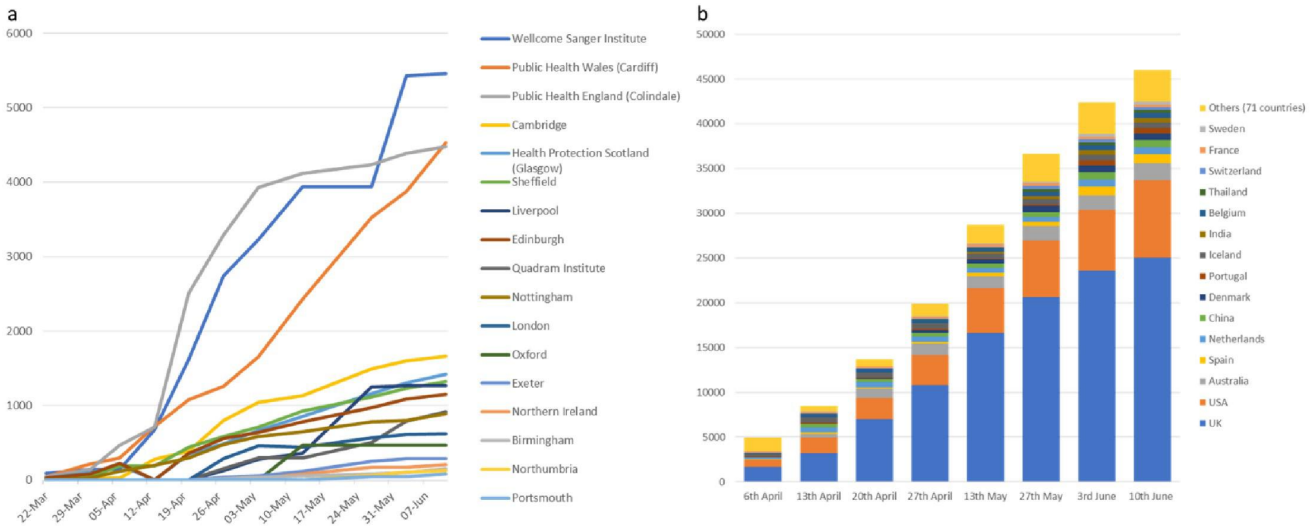

Figure 1: **a)** Number of SARS-CoV-2 genomes sequenced and analysed by the COG-UK centres by date. A total of 25,052 genomes have been sequenced across all COG-UK centres. Any week-to-week decreases owe to increased stringency in quality control parameters affecting inclusion of genomes. **b)** Number of SARS-CoV-2 genomes sequences reported (in MRC CLIMB and GISAID). Data shown up to 10th June.

Highlighted findings with public health implications

- The UK experienced a high volume of SARS-CoV-2 introductions, mainly from European countries, followed by local transmission within the UK. Analyses using large scale genomic surveillance coupled with data on the source/volume of inbound travellers and estimates of worldwide infection rates could be used as a platform to inform evaluation of future trends in virus introduction and efforts to limit imports.
- The volume of introductions was such that the impact of any individual event (e.g. sports matches or conferences) on the number of cases introduced to the UK as a whole was likely negligible.
- Updated analysis of a survey of SARS-CoV-2 cases in six London care homes supports implications drawn after preliminary analyses that infection prevention (including screening and exclusion of SARS-CoV-2 positive staff) and control interventions to reduce transmission between residents are important. Further investigations to understand the patterns of spread through care home systems are needed and should include a focus on status as resident or staff, symptomatic or asymptomatic, and whether the staff member works in more than one care home.

Analysis updates

Preliminary analysis of SARS-CoV-2 importation & establishment in the UK

https://virological.org/t/507

Study leads

Oliver Pybus (University of Oxford) and Andrew Rambaut (University of Edinburgh)

Question addressed

What were the trends in number and sources of introductions of SARS-CoV-2 lineages into the UK during the early months of the outbreak?

Methodology

Combined data on the number of inbound travellers to the UK, estimated number of infections worldwide and COG-UK SARS-CoV-2 genome sequence data. See above link for detailed methods.

Findings

To date, the UK epidemic comprises at least 1356 independently-introduced UK transmission lineages (i.e. two or more cases descending from a shared single introduction, followed by transmission within the UK). These transmission lineages occur owing to inbound international travel. However, this number is expected to be an under-estimate.

The proportion of UK transmission lineages newly detected for the first time decreased to negligible levels by early May 2020 and many UK transmission lineages now appear to be very rare or extinct, as they have not been detected by genome sequencing for >4 weeks.

Estimates of TMRCA (time of the most recent common ancestor) revealed that the majority of UK transmission lineages are dated to mid-to-late March, although this represents the date of first detection of a lineage, not necessarily the virus importation date.

Data on the number of inbound travellers were combined with estimates of SARS-CoV-2 cases worldwide. A period of high volume of inbound travel through March could be observed to coincide with growing numbers of active cases in other countries. From the beginning of April, a ~95% drop in inbound international travel to the UK coincided with a peak and then decline in the estimated number of cases worldwide.

Estimates of import intensity (an empirical estimate of the daily intensity of SARS-CoV-2 importation into the UK) were found to match, but precede, the pattern of TMRCAs of UK transmission lineages, with the difference between the curves representing the lag time between importation and first observation of the

transmission lineage in COG-UK genome data. Statistical modelling revealed the duration of the importation lag to be 10.7 days on average, although this number is expected to vary depending on the number of genomes within a lineage. Accordingly, lineages representing >15 genomes had a lag of just 4.2 days, which is the best current estimate of the duration between arrival of an infected passenger and first onward transmission event in the UK.

Combination of TMRCAs and the statistical lag model revealed that 80% of the importation events that gave rise to detectable UK transmission lineages occurred between 28th February and 29th of March 2020.

Estimates of the number of inbound travellers from those countries with both high numbers of inbound travellers and ongoing COVID-19 deaths between Jan-Apr 2020 allowed importation intensities to be calculated for each country (Figure 2). The highest importation intensities were seen for those countries where a window of time existed in which large numbers of inbound travellers coincided with high disease prevalence (e.g. Spain, France and other European countries). Note that in this analysis, assignment of lineages to countries of origin is based on phylogenetic placement without supporting individual epidemiology (i.e. without known links to travel).

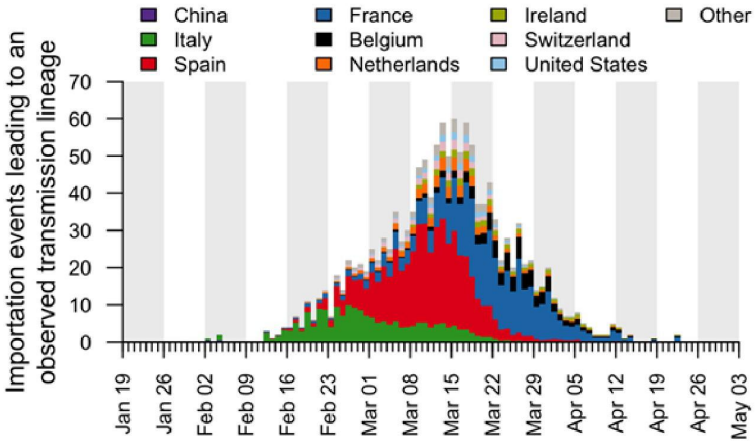

Figure 2: The estimated number of importation events that are attributable to inbound travellers from each of several source countries. Values shown are per day and not cumulative. Estimated dates of importations are obtained by combining the size-dependent importation lag model with the TMRCAs. Note that this is a statistical inference of the overall importation process, and cannot ascribe a specific source location to any given UK lineage.

Overall, ~34% of detected UK transmission lineages arrived through inbound travel from Spain, ~29% from France, ~14% from Italy and ~23% from other countries. The relative contribution of these sources were highly dynamic.

Key conclusions

The scale of COG-UK genomic data when combined with information on inbound traveller information and estimates of the number of global cases has revealed a high frequency of virus imports that led to onward viral transmission in the UK.

Contrary to media coverage which focussed on the earliest importation events from China and East and Southeast Asia, importations from these locations constitute a tiny fraction of all import events that resulted in detectable UK transmission lineages.

The high volume of inbound travel from countries with high COVID-19 case numbers indicates that individual events likely made a negligible contribution to the overall number of imports at that time. Large-scale and longer-term trends in prevalence and mobility are much more important.

Limitations

This is a preliminary analysis using a newly developed analytical framework, and the estimates generated do not capture all statistical uncertainty involved.

This work does not attempt to measure the relative contributions to the UK epidemic of importation versus local transmission, nor model the possible impact of public health interventions on virus introduction.

Proposed next steps

This work provides a basis for evaluating future trends in virus introduction. Viral introduction and transmission dynamics could be taken into account when planning and modelling future public health actions in the context of international travel.

The relative contributions of SARS-CoV-2 importation and local transmission to early epidemic growth in all countries warrants further investigation, once sufficient genome data is available.

SARS-CoV-2 genomic epidemiology case study - Scotland

https://doi.org/10.1101/2020.06.08.20124834

Study leads

Emma Thomson (University of Glasgow), Matt Holden (PHS) and Kate Templeton (Royal Infirmary of Edinburgh)

Question addressed

What insights can genomic epidemiological surveillance provide about the first four weeks of emergence of SARS-CoV-2 in Scotland, from the first detected case on 1st March 2020 through to 1st April 2020?

Methodology

Genomic epidemiology approach based on combination of genome sequence data from individuals with a laboratory-confirmed diagnosis of COVID-19 with phylogenetic and epidemiological analysis. See the above link for detailed methodology.

Findings

Complete SARS-CoV-2 genomes were sequenced from 452 individuals with a confirmed diagnosis, accounting for ~20% of the 2310 confirmed cases in Scotland for the period. Of the 452 individuals, 60% reported no travel, 26% had no travel history recorded, and 14% reported travel outside of Scotland, only one of which was internal to the UK; 57 were visits to various European countries (predominantly Italy) and 3 were individuals who had returned from a cruise holiday in the Caribbean.

The viral genomes exhibited limited variability overall, with an average of 3.4 non-synonymous and 1.8 synonymous substitutions compared to the original Wuhan-Hu-1 genome. The majority of viruses belonged to the global B lineage (432/452) and common amino acid replacements encoding D614G and P323L in the spike and nsp12 proteins, respectively, could be observed. A combined phylogenetic and epidemiological analysis suggests that there were between 113 and 276 separate introductions of SARS-CoV-2 into Scotland. Viral genomes were closely related to those circulating in other European countries (including Italy, Austria and Spain) and the introductions predate travel restrictions. The majority of introductions were single cases that were not linked with further cases over time. However, 48 introductions did result in case clusters consisting of more than two individuals and were associated with transmission in varied settings including care homes, community, and a conference held in Edinburgh in late February.

The first case not associated with travel was detected three days after the first confirmed case in Scotland, indicating that introductions were occurring undetected prior to the first known case. A notable shift from travel-associated introduction to sustained community transmission was apparent in multiple clusters after 11 days, coincident with an increase in the median age of individuals infected from 44 years old in the first week to 62 years old in the fourth week. This shift to community transmission preceded the introduction of 'lockdown' countermeasures on 23rd March.

Comparison of the genome of a virus sampled from a HCW with virus from patients in the hospital ward in which they were working revealed that they belonged to distinct lineages, indicating community transmission rather than HAI in this case.

Updated analysis of SARS-CoV-2 survey of 6 London care homes

Study leads

Richard Myers, Natalie Groves, Ulf Schaefer (Public Health England)

Analysis details

This is an update on a survey of SARS-CoV-2 infections among staff and residents from six London care homes (A-F). A preliminary analysis and background was included in COG-UK Report #5 (7th May 2020).

Findings

Staff and residents from six care homes in London were tested for SARS-CoV-2 infection and the samples from 158 individuals who were PCR positive for infection were then used for whole genome sequencing analysis. Of these, 99 samples, distributed amongst all the care homes, yielded genomes of sufficient quality for analysis; 31/99 were from staff and 68/99 were from residents. Sequences were aligned using mafft (version 7.310), manually curated and a phylogenetic tree was built using IQtree (version 2.04). The phylogenetic tree (Figure 3) was coloured to indicate care home of origin and annotated to indicate sequences derived from staff members and sequences from residents who had died. In order to place care home derived sequences within a comprehensive background of SARS-CoV2 genomes from within the UK, the care home sequences from this study were identified within the COG consortium maximum likelihood phylogeny containing 27768 sequences.

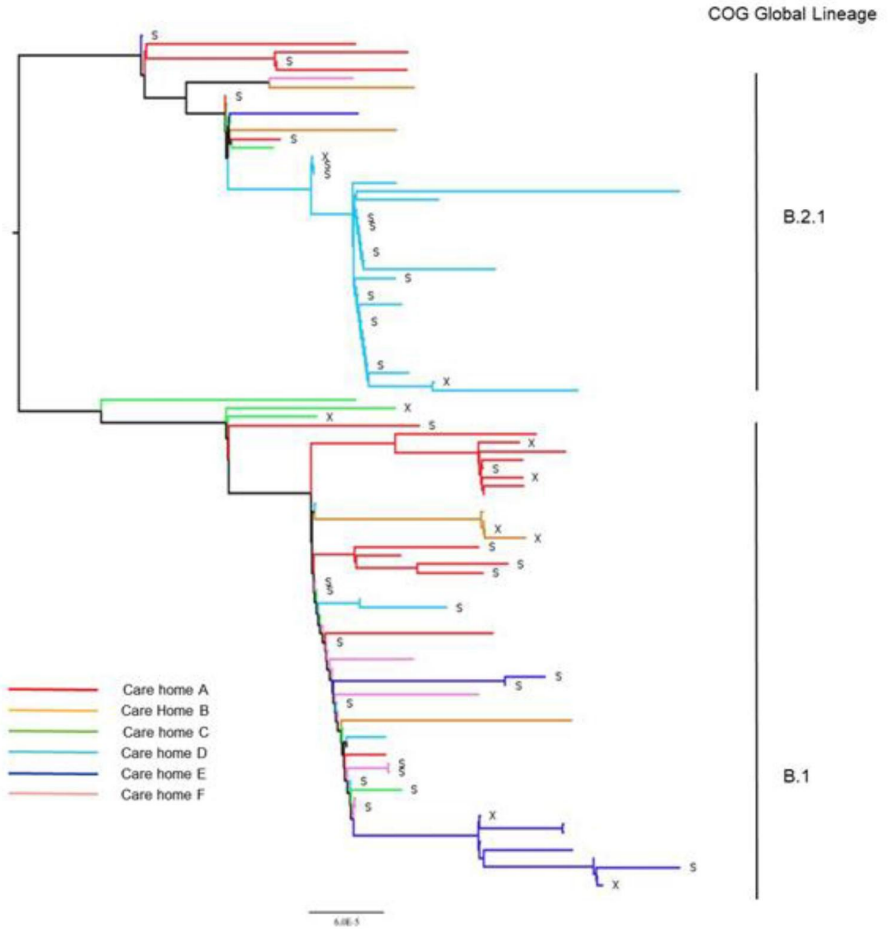

Figure 3. Maximum Likelihood phylogeny of 99 SARS-CoV-2 genomes from individuals within six care homes. Coloured branches are used to indicate the care home, staff are annotated on the tree with (S), genomes from patients who died after testing positive for covid-19 are shown with (X). Unannotated tips in the phylogeny represent genomes from care home residents.

Phylogenetic analysis indicated the presence of clusters from care homes A, B, D, E present in both the phylogeny from care home sequences (Figure 3) and within the large background dataset (Appendix, Figure S1). The largest cluster (care home D) contained 28 sequences of which 15 sequences exhibited zero SNPs difference and the maximum distance between sequences was three SNPs. The presence of clusters containing care home sequences, that did not contain background sequences and were distinct from that background, provided evidence for introduction and subsequent spread of a SARS-CoV2 strain in the care home setting.

Each of the six care homes contained SARS-CoV-2 genomes from lineages B.1 and B.2 and the distance between sequences in the large cluster (n. 28) in care home D (lineage B.2.1) and the sequences in lineage B.1 were 13-18 SNPs. This provides evidence for multiple introductions of the virus into care home settings. The placement of sequences in the phylogeny indicated that care home A exhibited three distinct sequence clusters along with six singletons, potentially representing up to nine separate introductions.

There were ten sequences that had a 0 SNP distance between them which were from three different care homes. However, these sequences were part of a large clade of sequences within the B.1 lineage (n. > 5,500). Comparison of these sequences with the background data showed that the care home sequences did not form a discrete cluster (Appendix, Figure S2). Some lineage B.1 sequences that were not from care homes were also identical to the ten sequences from the three different care homes. It is therefore possible that identical viruses were introduced from other settings into all three homes separately, instead of being transferred from home to home. This observation indicates that genomics can neither exclude nor confirm that the cases in separate homes were linked.

All care home clusters of SARS-CoV-2 genomes included at least one staff member, apart from those from the care home with no PCR positive staff. Other than this observation, there was no genetic signal within the SARS-CoV-2 genomes that differentiated staff and residents or symptomatic and asymptomatic individuals. The ten available sequences from individuals who died were distributed across the diversity of sequences derived from the care homes (Figure 3) and were closely matched to sequences derived from non-fatal cases in the same locations, indicating the absence of a particular strain associated with deaths in this study.

Appendix 1

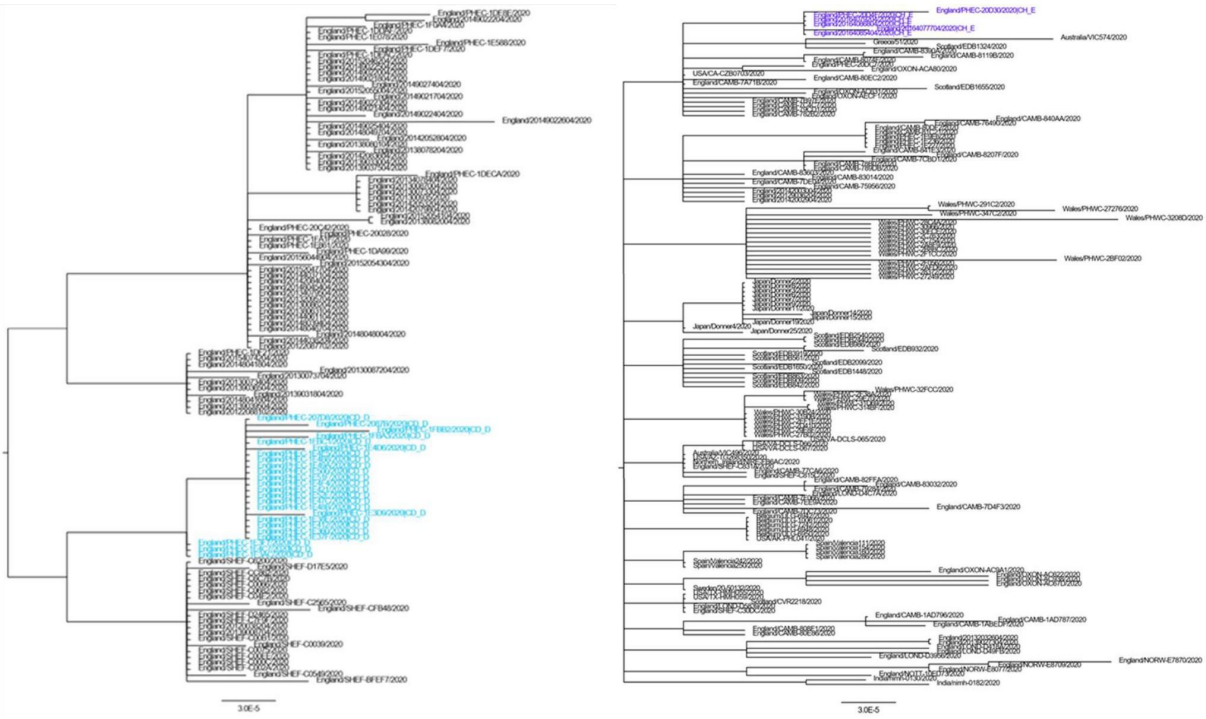

Figure S1 | Image taken from COG Consortium phylogeny of 27768 SARS-CoV-2 genomes. The taxa labelled in light blue are a cluster of sequences from Care home D, The cluster of taxa in dark blue are sequences from Care home E. In both examples the cluster of sequences derived from care home settings is retained in the presence of a large background

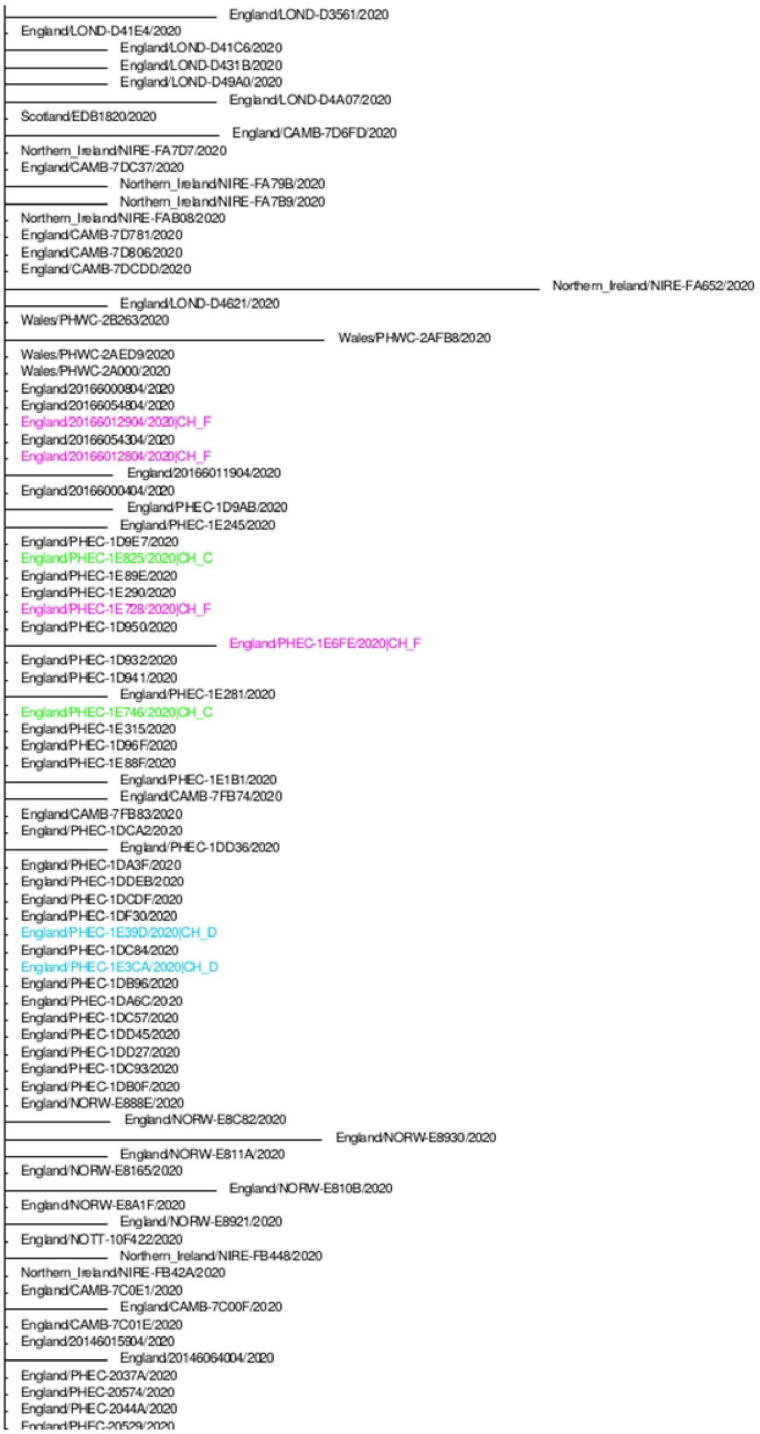

Figure S2 | Image taken from COG Consortium phylogeny of 27768 SARS-CoV-2 genomes. Coloured taxa are used to illustrate the location of sequences derived from care home settings. Seven of the eight coloured taxa are identical (two additional sequences are not shown in this portion of the phylogeny). These sequences are part of a large lineage of SARS-Cov-2 genomes (>5,500) with little sequence diversity. Sequences shown within this portion of the image cannot be considered as part of a cluster of care home cases. .