COVID-19 Genomics UK (COG-UK) Consortium

Report #5 - 7th May 2020

Executive Summary

- A further 4 sequencing centres have joined COG-UK. bringing the total to 14 including the Wellcome Sanger Institute (WSI). The number of SARS-CoV-2 genomes sequenced and analysed to date is 10,843. The UK has reported the largest number of genomes of any individual country in the pandemic to date, accounting for over half of the global total.
- Data sharing agreements having been formalised and metadata from patient electronic health records is being made available to COG-UK for integration into analyses.
- Samples and associated metadata from Lighthouse national testing centres are now being received, sequenced and analysed.
- Initial analyses on global viral lineages indicate that SARS-CoV-2 lineage diversity has been decreasing in the UK since mid-March, likely owing to extinction of some viral lineages and rapid transmission increasing the frequency of others.
- A preliminary genomic survey in five care homes in London suggests a pattern of multiple SARS-CoV-2 introductions followed by within care home transmission (and potential transmission between different care homes).
- SARS-CoV-2 genome variation is not likely to be a major confounder for diagnostic testing at present since the majority of primer/probe regions known to be in current use exhibit low levels of variation.

COG-UK update

In the past week, four additional sequencing centres have joined COG-UK (University of Birmingham, University of Liverpool, University College London and the Quadram Institute in Norwich). This total number of active COG-UK sites is now 14, including the WSI. A further two sequencing centres are expected to join us in due course (Figure 1).

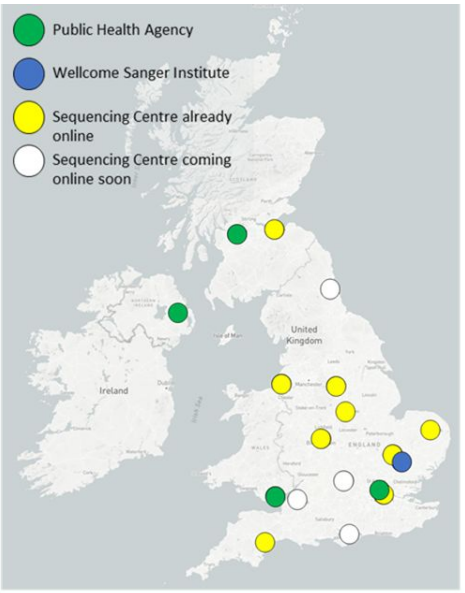

Figure 1: Location of current COG-UK sites.

By the data cut-off for this report, the total number of viral genomes is 10,843 (Table 1).

Since its inception, COG-UK has generated more than three times the number of genomes reported by any other single country,

accounting for more than half of the total number of SARS-CoV-2 genomes reported globally (Table 2; Figure 1).

Data sharing agreements are now in place and the first batch of metadata from 3000 patient electronic health records is being made available to COG-UK for integration into analyses. The volume of available patient data will increase substantially in the coming weeks, which will enable the use of SARS-CoV-2 genome data to inform public health questions.

Samples and associated metadata from two Lighthouse national testing centres are now being received, sequenced and analysed at the WSI. Both the Milton Keynes and Alderley Park testing centres are providing samples from Lighthouse labs to be sequenced weekly.

On the 24th of April a letter was sent from PHE to NHS laboratories introducing COG-UK and asking for the contribution of samples. To date, 20 laboratories have approached WSI and the necessary arrangements are now being put in place to receive samples.

COG-UK has an open community ethos, ensuring that data is updated and made publicly available via the consortium website (https://www.cogconsortium.uk/data/).

Sequencing Centre	Number of SARS-CoV-2 genomes					
	22/03	31/03	06/04	13/04	20/04	27/04
Birmingham	-	-	-	-	-	22
Cambridge	-	33	33	283	392	801
Edinburgh	32	86	223	220*	360	558
Exeter	-	-	4	4	4	35
Glasgow	27	102	203	195*	356	490
Liverpool	-	-	-	-	-	124
London	-	-	-	-	-	290
Nottingham	1	28	116	202	299	483
Public Health England (Colindale)	44	147	469	713	2511	3294
Public Health Wales (Cardiff)	48	218	298	719	1077	1262
Quadram Institute	-	-	-	-	-	158
Wellcome Sanger Institute**	95	143	143	676	1622	2741
Sheffield	13	49	190	190	440	585
TOTAL	**260**	**806**	**1679**	**3202**	**7061**	**10843**

Table 1: Number of SARS-CoV-2 genomes sequenced and analysed by the COG-UK centres. *Some figures are lower than previous week owing to increased stringency in quality control parameters affecting inclusion of genomes. **Samples from multiple sites (including Cambridge, Northern Ireland and Bristol) have been sequenced at Wellcome Sanger Institute.

Country	Number of SARS-CoV-2 genomes						
	11/03	22/03	31/03	06/04	13/04	20/04	
UK	24	260	806	1679	3202	7061	10843
USA	70	182	744	793	1795	2321	3320
Australia	18	30	71	385	391	1067	1305
Iceland	-	-	343	343	601	581	581
Netherlands	28	190	190	190	585	561	559
China	206	280	296	344	394	374	428
Belgium	-	-	-	255	342	388	427
Denmark	-	-	-	9	9	9	337
Luxembourg	-	-	-	56	86	86	232
France	10	61	119	167	205	221	223
Spain	-	3	42	89	151	148	191
Russia	-	-	-	3	4	37	150
Canada	3	17	129	129	130	122	137
Hong Kong	-	-	-	89	90	122	117
Japan	10	12	84	101	103	102	101

Table 2: Number of SARS-CoV-2 genomes sequences reported in GISAID. Fifteen highest countries shown.

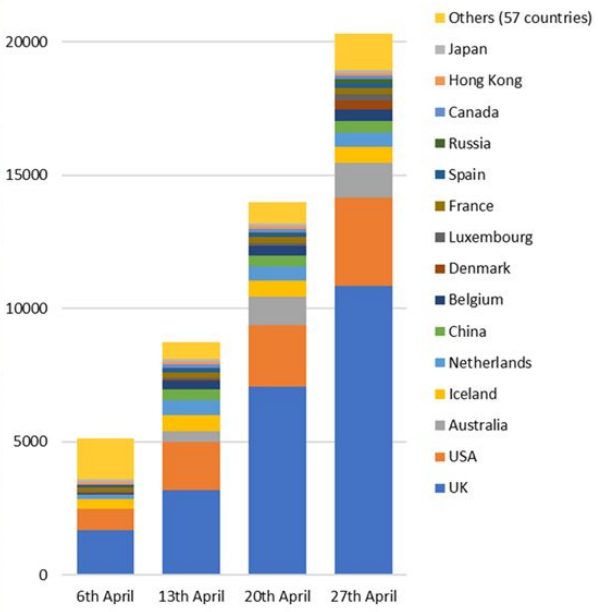

Figure 2: Number of SARS-CoV-2 genomes sequences reported in GISAID.

Highlighted findings with public health implications

- Preliminary analyses of SARS-CoV-2 genomes originating from people in care homes demonstrate multiple introductions into single care homes, followed in some cases by transmission to other residents. Sequences were generated from symptomatic and asymptomatic care home staff members. This indicates a need for infection prevention and control interventions to reduce transmission between residents, and a focus on staff, which includes screening and exclusion of COVID-19 positive staff and review of working practices across more than one care home. It is essential that further studies include metadata on status as resident or staff, symptomatic or asymptomatic, and whether the staff member works in more than one care home. This will enable a better understanding of the patterns of SARS-CoV-2 spread through care systems.

Analysis updates

Current population structure of SARS-CoV-2 in UK

For updated views of the geographic distribution of main SARS-CoV-2 lineages in the UK, see Appendix 1.

Rise and fall in the diversity of SARS-CoV-2 lineages in the UK
Oliver Pybus, David Aanensen, Andrew Rambaut

SARS-CoV-2 infections worldwide are classified into a number of "lineages" according to differences in the virus genome (1). Although there is no evidence these lineages have different biological properties, they are useful in tracking the number and size of different chains of transmission.

Figure 3 shows the diversity of global lineages detected by COG-UK in England, Scotland, and Wales during 2020. There are insufficient genomes from Northern Ireland for comparison. The diversity score is high when there are many lineages of equal frequency, and low when most infections are caused by just a few, dominant lineages.

The diversity of SARS-CoV-2 lineages in the UK initially rose and then started to decline around mid-March. The rise in diversity is due to the importation of different lineages into the UK by international travellers, and coincides with the period before international travel into the UK dramatically fell. Diversity of lineages rose earlier in England, perhaps reflecting a wider range of origin countries for air passengers to London.

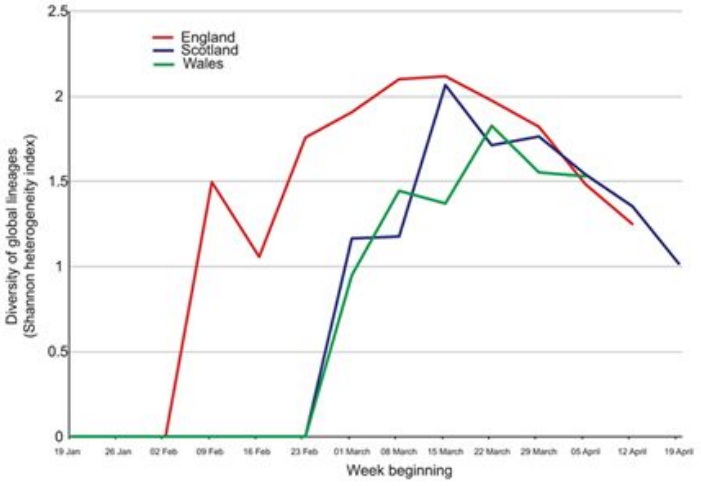

Figure 3: Diversity of global SARS-CoV-2 lineages circulating in regions of the UK over time.

The fall in diversity has two causes. First, some of the lineages introduced into the UK will not have become established and likely became extinguished during the lockdown (which began on 23rd March). Second, a few lineages have transmitted rapidly and therefore grew in relative frequency, resulting in the epidemic becoming more dominated by a smaller number of lineages. This view is reinforced by noting that the decline in lineage diversity began when the number of UK cases was still rising rapidly (diversity is thus distinct from the number of cases through time, a dynamic that only genome sequencing can reveal). The rate of diversity decline is similar for all three nations, indicating that the growth dynamics of the epidemic in each are similar.

Further analysis using UK lineage information and linking with postcode data will provide a high resolution view of lineage diversity in the UK. We expect lineage diversity to decrease to a point where we can distinguish dominant local lineages and track viral spread from one part of the UK to another.

(1) https://www.biorxiv.org/content/10.1101/2020.04.17.046086v1.full

Preliminary analysis of SARS-CoV-2 genomes from London care homes

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

Between 13th and 15th of April, 210 samples were taken from staff and residents from five London care homes (here identified as Alpha, Bravo, Charlie, Delta and Echo). Of these samples, 91 were positive for SARS-CoV-2.

Whole genome sequences of sufficient quality were generated for 55 of these samples and used for phylogenetic analysis within each care home, and for comparison with 277 non-care home SARS-CoV-2 genomes originating in the greater London area.

Care home Bravo contained the largest number of genomes analysed (27), exhibiting two distinct groupings. One cluster consisted of 20 closely related SARS-CoV-2 genomes, with only a few single nucleotide polymorphisms (SNPs) different among cluster members. The second cluster consisted of 7 genomes that were less closely related, with on average 12-15 SNPs different from samples in the larger cluster. These findings suggest multiple SARS-CoV-2 introductions into care home Bravo, one that led to the larger cluster, and at least one other introduction, likely several, that led to the smaller cluster.

A similar pattern of introduction of genetically distinct viral lineages was observed in care homes Alpha and Delta, although less data was available for these sites. Insufficient data was available for care homes Charlie and Echo to draw any conclusions.

Comparison with non-care home SARS-CoV-2 genomes from the Greater London area revealed no large-scale clustering of viruses from care home settings, indicating multiple introductions from the wider community (rather than a limited number of introductions followed by wider spread within the care home setting). In particular for care home Bravo, comparison with non-care home genomes indicated at least four separate introductions in total.

Introductions often included a staff member, but no common pattern for staff/resident or symptomatic/asymptomatic annotation could be detected.

Putative clusters were identified containing samples from multiple care homes suggesting transmission of SARS-CoV-2 between care homes. Not enough data was collected to determine whether this related to staff movement between care homes or residents mixing in a common setting.

Further efforts to analyse SARS-CoV-2 genomes from care home settings and the surrounding area will enable common routes of introduction and transmission to be identified and identify potential opportunities to limit spread through ensuring infection control measures are as effective as possible.

For the full report, see Appendix 2

SARS-CoV-2 variation affecting diagnostics and therapeutics

a) **Weekly primer analysis:** Richard Myers, Eileen Gallagher, Natalie Groves, David Williams (Public Health England)

The major obstacle to performing primer/probe analysis is a lack of information about the specific assays that are being used within the UK. To that end PHE is:

- Working with PHE colleagues involved in the roll out of testing to laboratories
- Utilising PHE clinical staff to work with networks within NHS to generate information

PHE is in the process of gathering lists of primer/probe sets that will start to be included in the next report.

We have generated a comprehensive analysis of variation in the predicted primer and probe binding regions for eleven of the assays that have been described publicly. This has been performed at a global lineage specific level (as designated by COG-UK), but not at a UK lineage level.

A summary has been generated (Table 3). This table is a count of the total number of mismatches to each primer and probe sequence in the COG-UK analysis. In this analysis 15,293 sequences were compared to the eleven sets of primer and probe sites. For the RdRP assay results are shown with and without the S (G/C) mismatch, present in every sequence.

The China CDC Orf1ab assay (used within PHE Colindale) has low level variation in the primer sites across a range of B lineages (B.1, B.2, B2.1) which are large lineages (>700 sequences) and B.1.18 smaller lineage (15 sequences). The probe site has one variant (T26C), which is present in >67% of one smaller lineage (B.1.15, 12 sequences) and appears in small proportions within other lineages.

The majority of predicted primer and probe binding regions have low level variation, with the exception of the RdRP assay.

Primer	F	R	P	Total
E	3	3	21	27
IP2	1	30	5	36
Orf1b	34	17	16	67
IP4	4	52	25	81
HKU N	36	34	11	81
N2	33	13	26	72
China CDC Orf1ab	10	31	57	98
N1	14	112	159	285
N3	188	59	61	308
China CDC N	12170	39	3	12212
RdRP (P2) (w/ S)	35	15293	8	15336
RdRP (P2) (w/out S)	35	0	8	43
RdRP (P1) (w/out S)	35	0	30587	31998
RdRP (P1) (w/ S)	35	15293	30587	45915

Table 3: Count of the total number of mismatches to each primer and probe in the COG analysis. F (forward primer), R (reverse primer), P (probe).

Appendix 1

Figure S1 | Latest geographic distribution of main SARS-CoV-2 lineages in the UK visualised using Microreact.

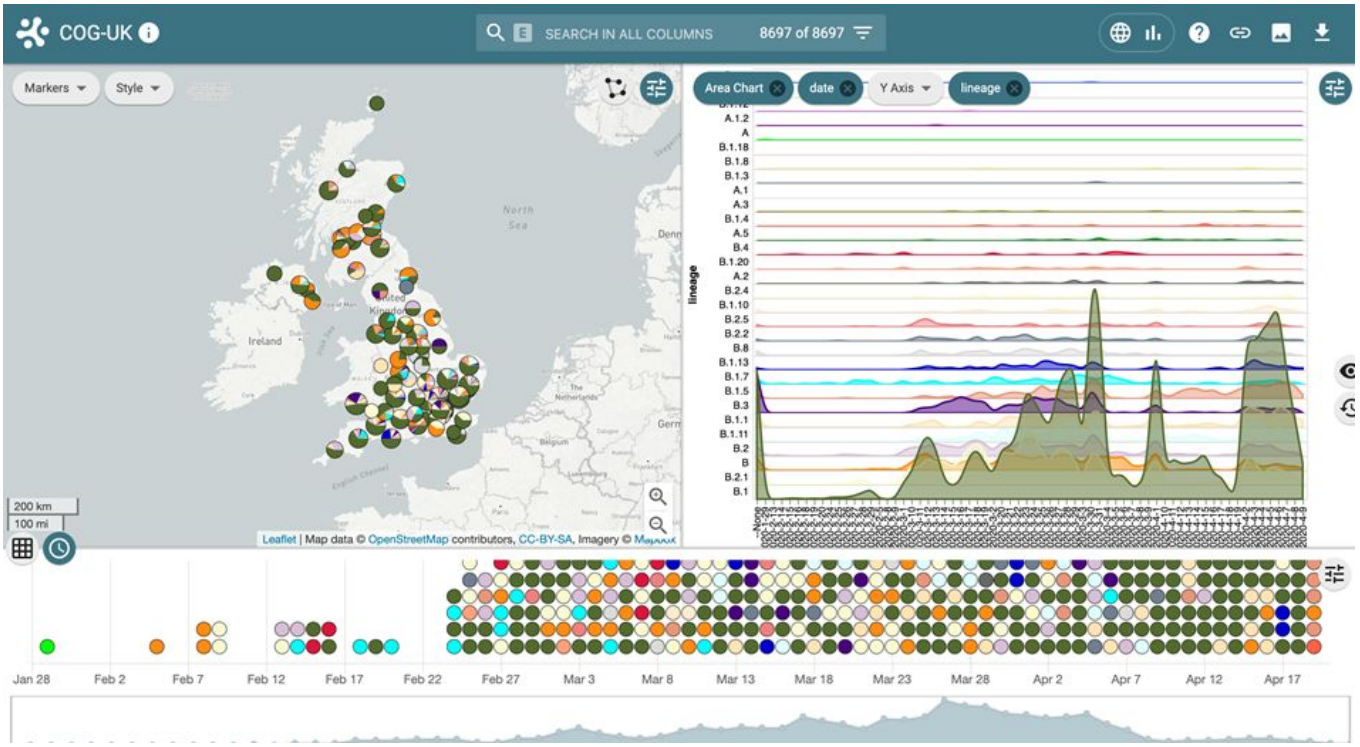

Figure S2 | Recent view of data visualised using Microreact. Upper left panel displays geographic distribution of main SARS-CoV-2 lineages in the UK. Upper right panel displays lineage frequency over time. Lower panels display a timeline of the sampling for viral genomes (dots) and total sample numbers (grey graph). Live link to the view in the above screenshot: https://microreact.org/project/COGconsortium/61146402

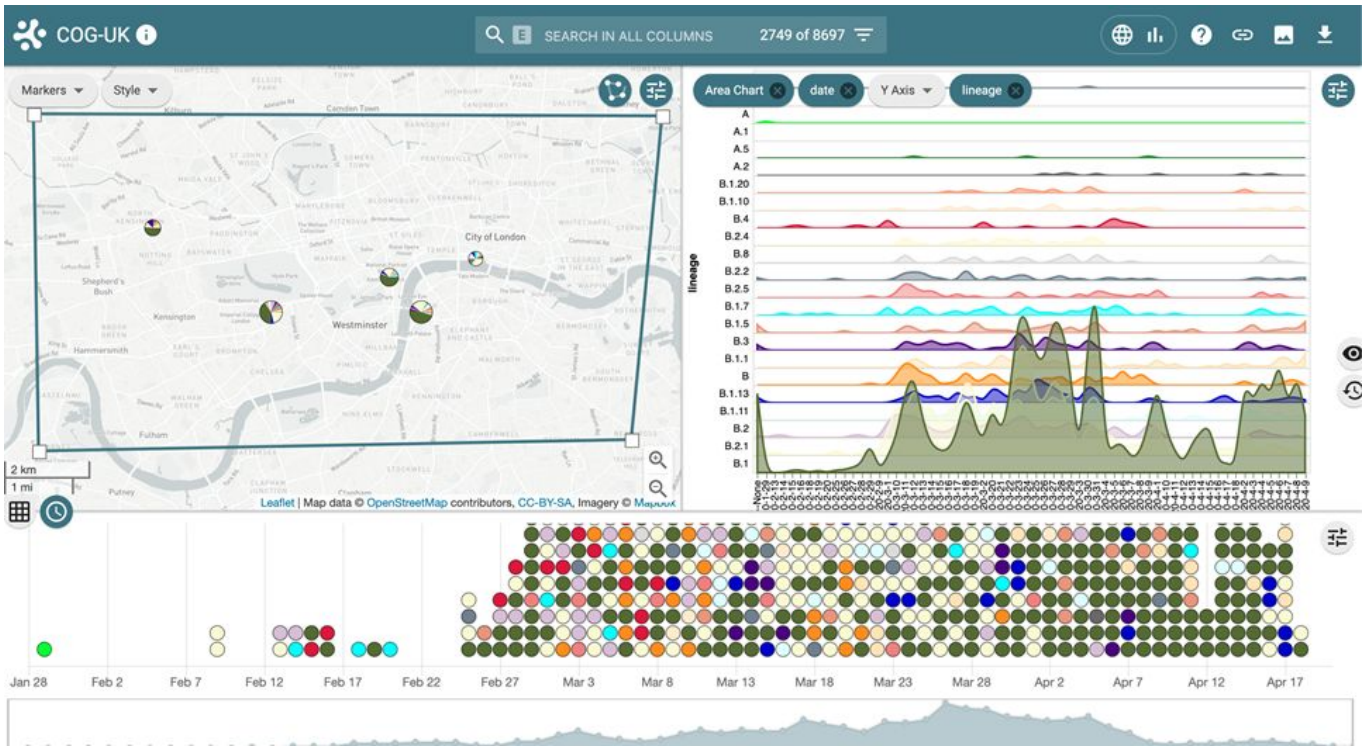

Figure S3 | Recent view of data visualised using Microreact. Upper left panel displays geographic distribution of main SARS-CoV-2 lineages in London. Upper right panel displays lineage frequency over time. Lower panels display a timeline of the sampling for viral genomes (dots) and total sample numbers (grey graph). Live link to the view in the above screenshot: https://microreact.org/project/COGconsortium/61146402

Appendix 2

Preliminary Analysis of SARS-CoV2 Genomes from London Care Homes (Public Health England)

Samples from staff and resident screening across five care homes were sequenced (n=75). Sequences for 16 positive care home samples were not yet available. These sequences were compared to other WGS sequences submitted to COG-UK and identified as originating in the Greater London area (n=277).

	Total number of samples w/ result	Number of positive samples	Number of samples sequenced	Bad quality sequence	Number of sequences analysed
Alpha	96 (S: 47, R: 49)	22 (S: 3, R: 18)	21	8	13
Bravo	69 (S: 31, R: 33, at home: 5)	44 (S: 14, R: 26, at home: 4)	35	8	27
Charlie	10 (S: 0, R: 9)	7 (S: 0, R: 7)	6	0	6
Delta	35 (S: 8, R: 27)	14 (S: 3, R: 11)	11	4	7
Echo	10 (S: 0, R: 10)	4 (S: 0, R: 4)	2	0	2

S – Staff, R - Resident

Whole genome sequencing was performed on samples using reverse transcription and PCR amplification of extracted viral RNA. Viral amplicons were sequenced using Illumina library preparation kits (Nextera) and sequenced on Illumina short read sequencing machines. Raw sequence data was trimmed and aligned against a SARS-CoV2 reference genome (NC_045512.2). A consensus sequence representing each base of the genome derived from the reference alignment. Consensus sequences were collated from each sample, assessed for quality and then aligned (mafft). Maximum likelihood phylogenetic trees were derived from sequence alignments using IQtree

The sequence data produced for four of the five care homes showed some samples with a high proportion of Ns. Those sequences that contained >30% undetermined bases were excluded from this analysis (n=20) leaving 55 sequences to be compared to the Greater London assemblies. These results are preliminary and conclusions derived from these analyses should be treated with caution.

Care Home Bravo

- Care home Bravo contained the largest number of genomes sequenced (27). The maximum likelihood phylogeny (Figure 1) for the samples from this care home showed two groupings of sequences; one cluster where the 18 of the 20 samples were 0 to 3 SNPs different from other cluster members (clade 1), a second smaller cluster (seven

samples) contained sequences that were less closely related to each other and were on average 12-15 SNPS different from samples in the larger cluster. The two samples in clade 1 that appear to be more distantly related contained variants that are likely to be due to poor sequencing quality as both these samples had higher proportions of N bases across the genome. Once these erroneous variants are excluded, both samples fall within the 3 SNP threshold observed across the other samples.

- Phylogenies were annotated to indicate whether the sample had originated from staff or residents and the symptomatic / asymptomatic status of each individual. None of the care home phylogenies indicated any pattern that could be detected using these annotations.

- The larger cluster exhibited a low sequence diversity pattern that would be expected if there was introduction and transmission of a single strain of SARS-CoV2 into the care home setting. The presence of a sequence outside of the main cluster indicated that there had most likely been at least one other introduction of SARS-CoV2 into care home Bravo.

Combining Genomes Derived from Care Homes with a Background of Genomes Derived from London Samples

Sequences derived from samples taken from the care homes (n. 55) were combined with 277 genomes from samples taken within Greater London. The phylogeny built from this larger dataset was used to compare the diversity of care home sequences in the context of genome diversity seen within London (Figure 2).

Analysis of this phylogeny indicated:

- Care home sequences were distributed across the phylogeny in multiple international lineages (B1, B2, B3). There was no large-scale clustering of sequences derived from a care home setting.

- The low diversity cluster of sequences derived from care home Bravo (clade 1) was maintained when the background sequences were included. This observation supported the initial conclusion of a single introduction and transmission event accounting for those 20 cases.

- The seven sequences that did not form part of the larger cluster within care home Bravo, were in a separate lineage to the more populated cluster (B1 not B2). These seven samples were also distributed across at least three distinct clusters of sequences from the background. This observation supports the conclusion that there was a separate introduction of SARS-CoV2 into care home Bravo and actually suggests that there are likely to have been at least four distinct introductions. All but one putative introduction in care home Bravo is a sample from a staff member, with the exception of the apparent multi-care home cluster (detailed below) which includes a staff member from a different care home (Alpha)

- A similar pattern of multiple introductions of genetically distinct strains was also observed in the cases of care homes Alpha and Delta, though there is less data available from these locations.

- Clusters were also identified that contained samples from multiple care homes. One low diversity cluster (0-2 SNPS) contained six samples from four care homes and no background samples. This observation suggested that there was transmission of SARS-CoV2 between care homes. Five of the samples were from residents and one from a staff member, making it difficult to hypothesise if the cluster was the result of staff movement or residents being discharged from a common setting.

Figure 1. Maximum Likelihood phylogeny of 27 samples from care home Bravo

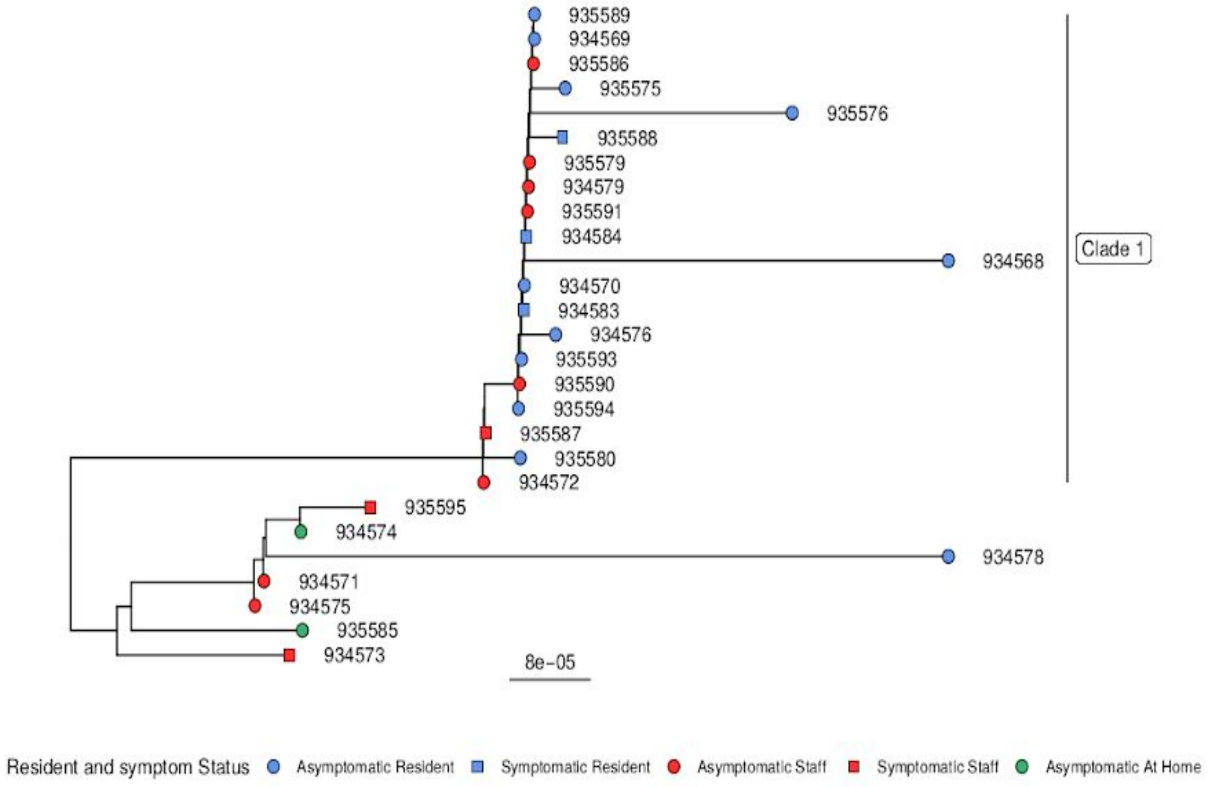

Figure 2. Maximum Likelihood phylogeny of SARS-CoV2 genomes. 55 genomes from five care homes – coloured boxes, 277 genomes from Greater London – black line