Lineages report for SHEF

This report gives summaries of UK specific lineages sequenced by SHEF for week 2020-09-13. There are time lags due to batching, curation and analysis, the most recently sampled sequence is 2020-08-25. The analysis (eg time since last sample) is therefore undertaken from this date. 2009 sequences in the UK from the sequencing centre SHEF have been included in this analysis.

A few notes: the size of a lineage may be due to a low amount of transmission of this lineage, but it is likely also that it just hasn't been sampled as frequently, especially for newer lineages. It's also important to realise that these lineages are *estimates* of how we think the virus is spreading in the UK after being introduced from abroad, as the low evolutionary rate of the virus makes it difficult to separate lineages with certainty.

The minimum number of introductions is 17 and the maximum is 719

Sequences which were replicates or too error-prone were removed from this analysis.

139 are lineages which only contained five sequences or fewer, and so have been left out of visualisation in the interests of clarity

Furthermore, those sequences which haven't been sampled in the last month are not shown.

Of the 3 that remain: 2 are pending extinction, ie last seen three weeks ago. 1 lineage has been continuously circulating.

The following table contains information about the ten largest lineages lineages and the number of sequences the dataset. Information about other lineages is found in the appendix, along with the raw data for all of the other figures.

Each entry is the count of sequences from each lineage in each country, with the percentage of the total sequences from that lineage that this count represents.

"Activity score" is calculated by taking the average gap between sampling for each lineage, and dividing it by the number of days since the lineage was last sampled. Therefore the higher the number, the more active the lineage is. If the score is above 1, then it has been sampled *more* recently than expected given its average gap size. We might interpret this as an increase in activity. If the score is below 1, it has been sampled *less* recently than expected given its average gap size, so we might interpret this as a decrease in activity.

The global lineages are correct as of the data release on 2020-07-20

It is written to "summary_files" as "lineage_summary.tsv" for further use, and the full list of lineages is available in the same directory as "all_lineages.csv"

Lineage name	England	Date range	Global lineage	Total
UK5	469 (100.0%)	Mar-02, Aug-25	B.1.1.10, B.1.1	469 taxa
UK1951	462 (100.0%)	Mar-18, Jul-23	B.1.1, B.1.1.1	462 taxa
UK1205	155 (100.0%)	Mar-13, Aug-10	B.1.1, B.1.1.1	155 taxa
UK319	91 (100.0%)	Mar-28, Jun-03	B.1, B.1.79	91 taxa
UK107	66 (100.0%)	Mar-05, May-03	B.2.1	66 taxa
UK51	51 (100.0%)	Mar-25, Jun-19	B.1, B.1.36	51 taxa
UK175	34 (100.0%)	Mar-20, Aug-03	B.1.8, B.1, B.1.11	34 taxa
UK2916	28 (100.0%)	Mar-03, May-28	B.1, B.1.98	28 taxa
UK267	27 (100.0%)	Mar-20, Apr-27	B.2	27 taxa
UK2028	25 (100.0%)	Mar-21, May-04	B.1.1	25 taxa

These data is represented in the figure one. Note that the number of sequences is likely to be due more to differing sampling efforts in different regions, rather than genuine differences in numbers of cases.

The raw data for this bar chart are in the table above.

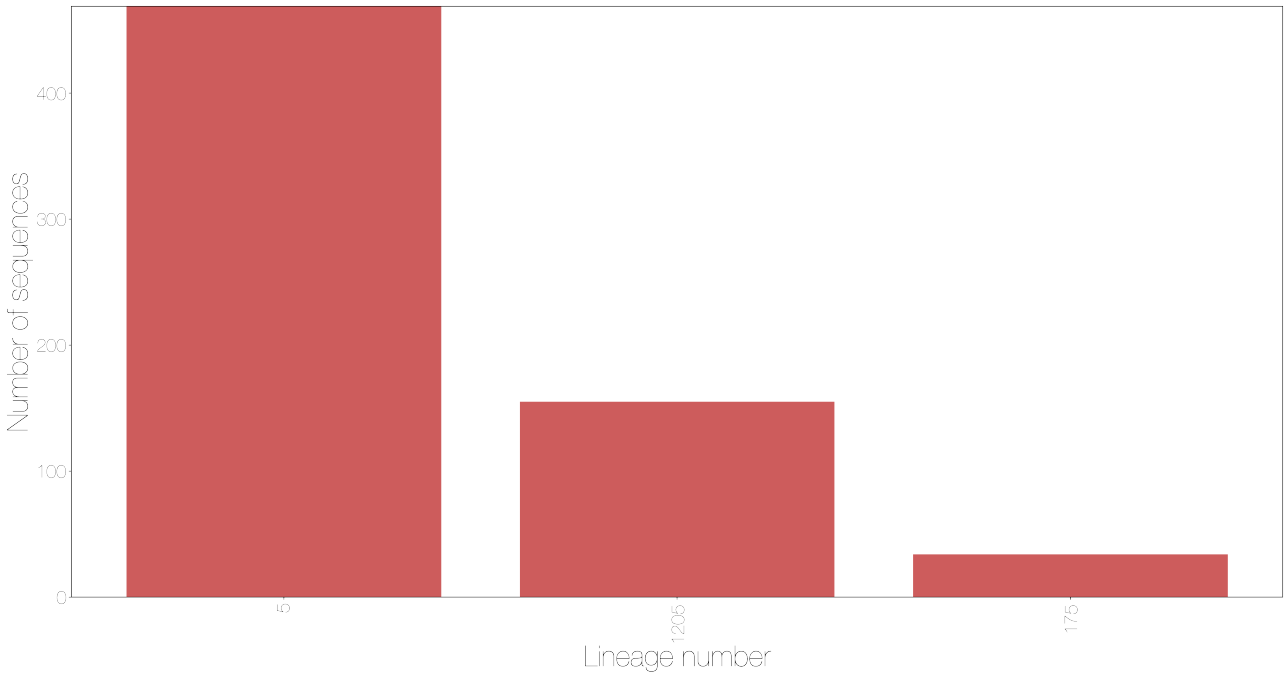

Figure 1: Number of sequences sampled in a lineage by country

Different sequencing centres have different delays in turn around from receipt of samples to submission of sequence data. This will affect all of the figures shown after this if lineages have geographical variation, as some regions have less up to date data.

```
---------------------------------------------------------------------------NameError
Traceback (most recent call last)<ipython-input-1-2620455843ef> in
<module>
      2     lag_dict, lags = dp.sequencing_centre_lags(taxa, sc_dict,
current_date, country)
      3 elif sequencing_centre != "":
----> 4     print("The lag for this sequencing centre is " +
str(lags[sequencing_centre]) + " days")
NameError: name 'lags' is not defined
```

The relative growth and decline of the ten most sampled lineages in terms of number of counties they are present in is shown in figure three.

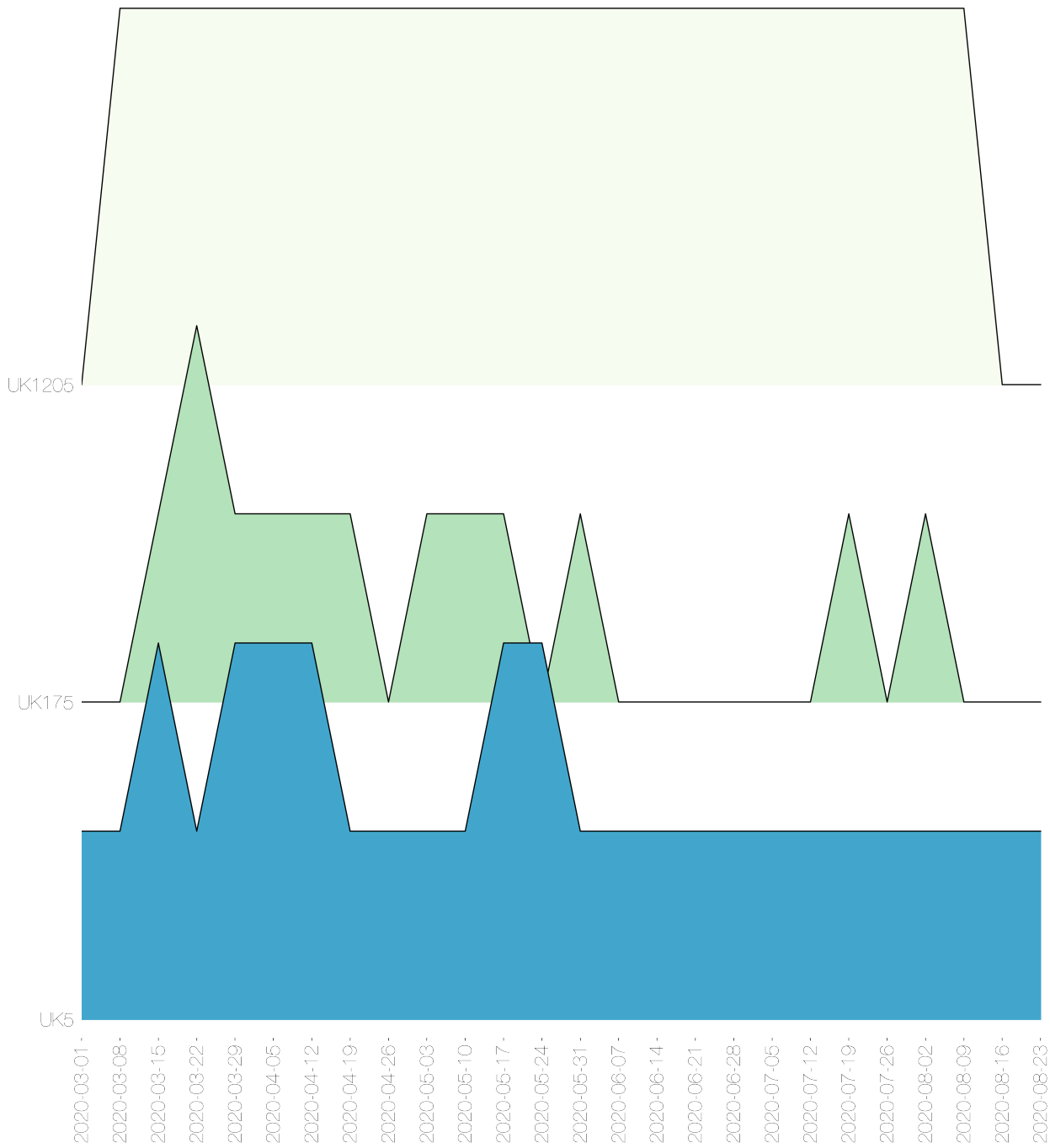

Figure 2: Lineages by number of adm2 regions present by epiweek

These lineages are shown on the timeline. Each line represents the length of the cluster, from oldest to most recent sampling date. The dots are sized by the number of sequences taken on that date, and again are colour coded by country. The raw data has been written to a summary file.

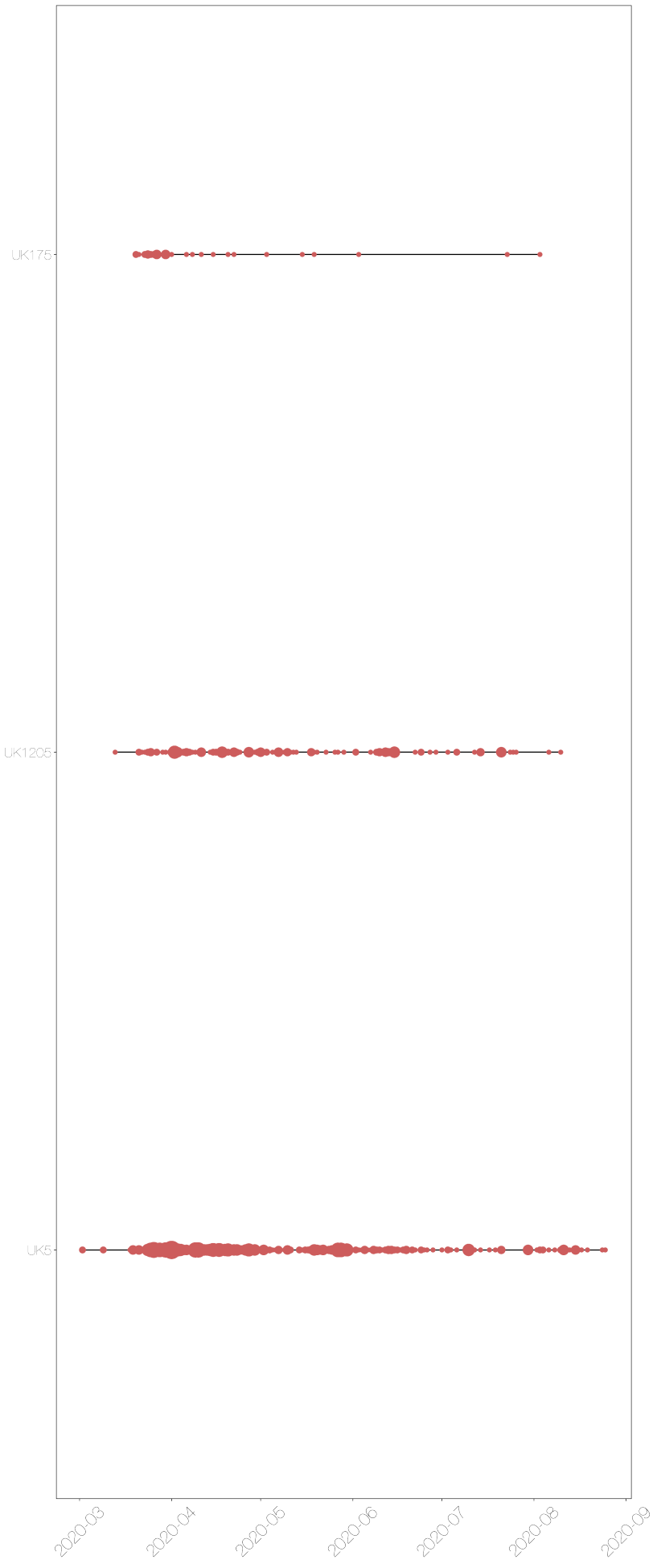

Figure 3: Timeline of lineages, sized by number of sequences from each country.

The date of first sequence in the cluster sampled by SHEF is shown in figure five for every cluster with date information.

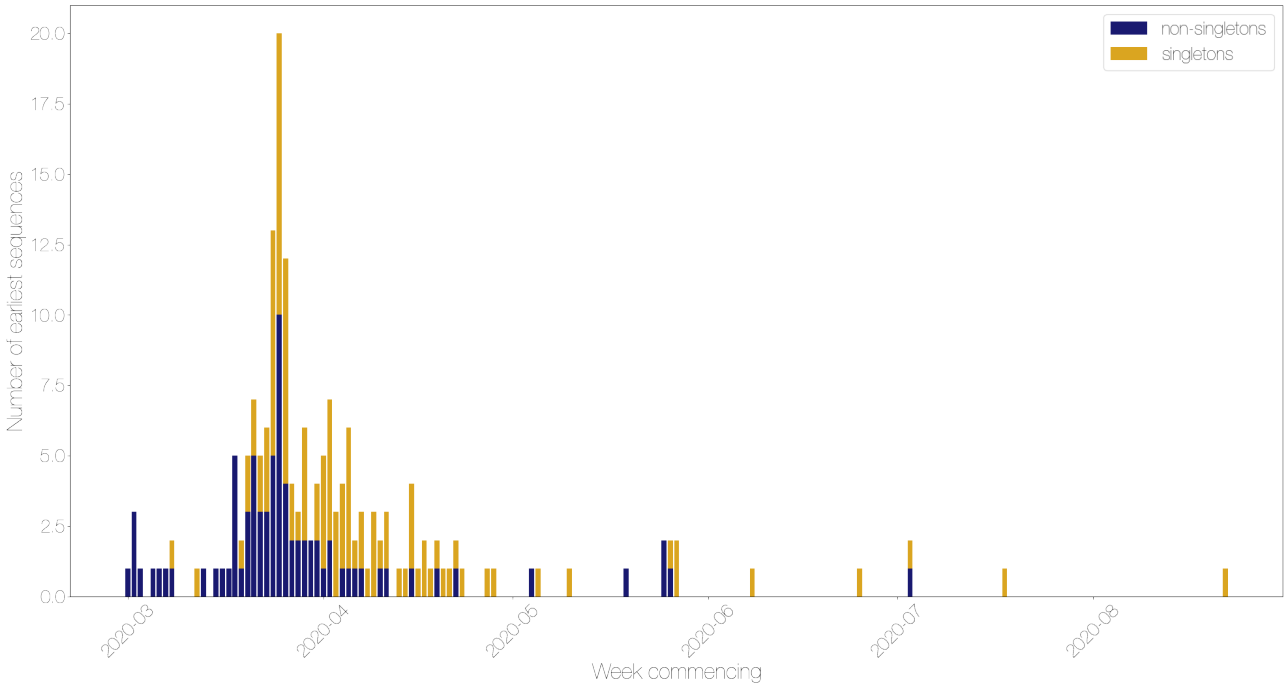

Figure 4: Lineage starts per week, split by singletons and non-singletons

For comparison, here is a plot of the day that every sequence was taken, coloured by country. Note that sequences without dates were not included.

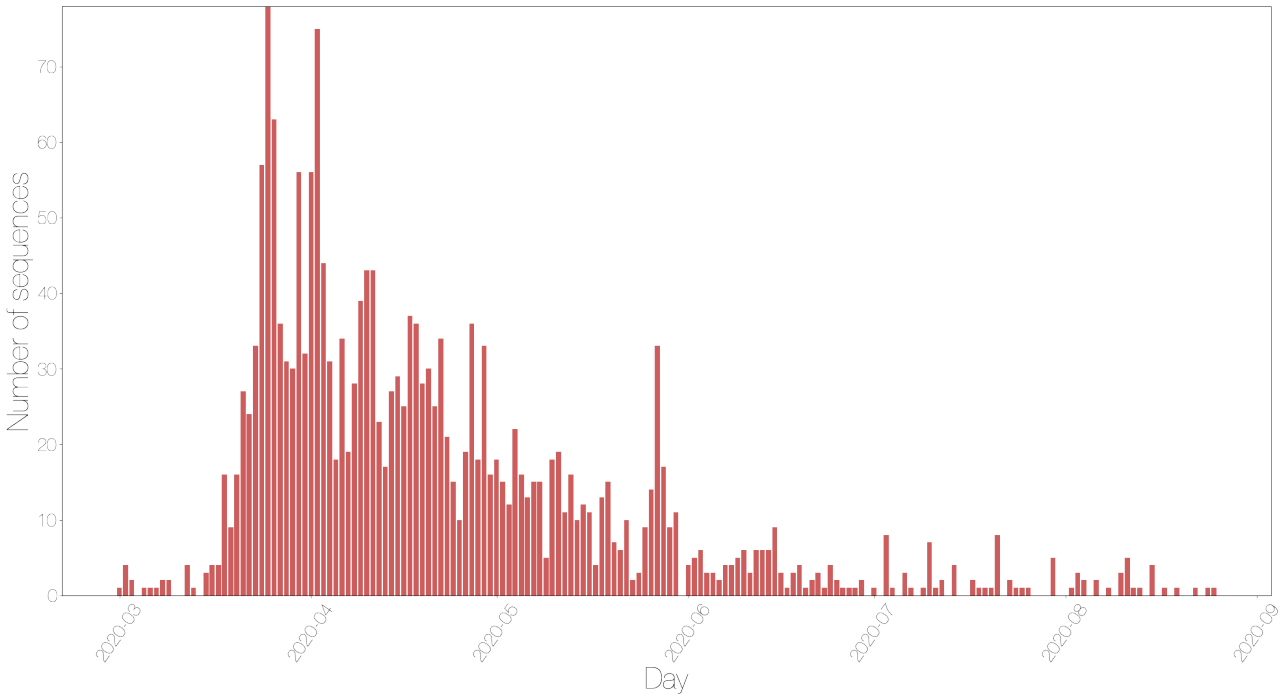

Figure 5: Sequences taken on each day by country

The map shows the number of sequences sampled in each admin2 region in the UK. The colour scale is the same for all four countries, but with different underlying base colours.

```
There are 4 sequences without enough geographical information to map
from this centre.
```

Figure 6: Map showing the number of sequences sampled by adm2 region

Other results modules for UK lineage analysis can be added in here if required.

Appendix

Below are the raw data tables for each of the figures in the report.

Table S1 Description of all lineages that have been circulating in the last month, and have more than 5 sequences.

Lineage name	England	Date range	Global lineage	Total
UK5	469 (100.0%)	Mar-02, Aug-25	B.1.1.10, B.1.1	469 taxa
UK1951	462 (100.0%)	Mar-18, Jul-23	B.1.1, B.1.1.1	462 taxa
UK1205	155 (100.0%)	Mar-13, Aug-10	B.1.1, B.1.1.1	155 taxa
UK319	91 (100.0%)	Mar-28, Jun-03	B.1, B.1.79	91 taxa
UK107	66 (100.0%)	Mar-05, May-03	B.2.1	66 taxa
UK51	51 (100.0%)	Mar-25, Jun-19	B.1, B.1.36	51 taxa
UK175	34 (100.0%)	Mar-20, Aug-03	B.1.8, B.1, B.1.11	34 taxa
UK2916	28 (100.0%)	Mar-03, May-28	B.1, B.1.98	28 taxa
UK267	27 (100.0%)	Mar-20, Apr-27	B.2	27 taxa
UK2028	25 (100.0%)	Mar-21, May-04	B.1.1	25 taxa
UK6	23 (100.0%)	Mar-22, Jun-16	B.1, B.1.75	23 taxa
UK1099	21 (100.0%)	Mar-24, May-14	B.1.1, B.1.1.5	21 taxa
UK199	19 (100.0%)	Mar-16, Jul-07	B.1.5, B.1	19 taxa
UK213	19 (100.0%)	Mar-21, Apr-10	B.2.1	19 taxa
UK501	19 (100.0%)	Mar-23, May-04	B.1	19 taxa
UK384	18 (100.0%)	Mar-15, Apr-02	B.2.1	18 taxa
UK23	18 (100.0%)	Mar-23, May-02	B.9	18 taxa
UK131	17 (100.0%)	Mar-20, Apr-10	B, B.15	17 taxa
UK607	16 (100.0%)	Mar-08, Apr-21	B	16 taxa
UK5741	15 (100.0%)	Mar-24, May-08	B.1	15 taxa
UK167	15 (100.0%)	Mar-25, May-01	B.1	15 taxa
UK55	14 (100.0%)	Mar-18, Apr-09	B.3	14 taxa
UK72	14 (100.0%)	Mar-01, Apr-09	B	14 taxa
UK945	12 (100.0%)	Mar-24, Jul-03	B.1.1	12 taxa
UK601	11 (100.0%)	Mar-22, Apr-22	B, B.10	11 taxa
UK2464	11 (100.0%)	Mar-28, Jun-10	B.1	11 taxa
UK12	10 (100.0%)	Mar-26, Apr-22	B.1.88	10 taxa
UK336	10 (100.0%)	Mar-24, May-07	B.1	10 taxa
UK1684	10 (100.0%)	Mar-31, May-30	B.1.1.1	10 taxa
UK600	9 (100.0%)	Mar-02, Apr-01	B.1.1	9 taxa
UK315	9 (100.0%)	Mar-06, Apr-06	B.2.2	9 taxa
UK1060	8 (100.0%)	Mar-26, Apr-29	B.1.1	8 taxa
UK809	8 (100.0%)	May-25, Jun-29	B.1.36	8 taxa
UK173	7 (100.0%)	Mar-22, Apr-06	B, B.21	7 taxa
UK5676	7 (100.0%)	Mar-21, Apr-03	B.2.4, B.2	7 taxa
UK4	7 (100.0%)	Mar-07, Apr-02	B	7 taxa
UK1076	7 (100.0%)	Mar-23, Apr-01	B.1.1	7 taxa
UK109	7 (100.0%)	Mar-27, May-10	B.1, B.1.99	7 taxa
UK311	6 (100.0%)	Mar-25, Apr-18	B.2	6 taxa
UK1946	6 (100.0%)	Apr-22, May-16	B.1.1	6 taxa
UK5084	6 (100.0%)	Mar-29, Apr-16	B.1	6 taxa
UK1278	6 (100.0%)	Apr-07, Apr-23	B.1.1	6 taxa
UK1683	6 (100.0%)	Mar-18, Apr-28	B.1.1, B.1.1.1	6 taxa
UK698	6 (100.0%)	May-19, Jun-23	B.1.1	6 taxa

Table S2 Raw data for figure two showing lags between the most recent sequence and current date for each sequencing centre

———————————————————————NameError Traceback (most recent call last) in 1 if not pillar2: —-> 2 lag_df = pd.DataFrame(lag_dict) 3 print(lag_df.to_markdown()) 4 else: 5 print("Table S2 is not appropriate for this report and so has been omitted.") NameError: name 'lag_dict' is not defined

Table S3 Raw data for figure three showing the number of admin2 regions a lineage is present in over time

Week commencing	UK5	UK1205	UK175
2020-03-01	1	0	0
2020-03-08	1	1	0
2020-03-15	2	1	1
2020-03-22	1	1	2

Week commencing	UK5	UK1205	UK175
2020-03-29	2	1	1
2020-04-05	2	1	1
2020-04-12	2	1	1
2020-04-19	1	1	1
2020-04-26	1	1	0
2020-05-03	1	1	1
2020-05-10	1	1	1
2020-05-17	2	1	1
2020-05-24	2	1	0
2020-05-31	1	1	1
2020-06-07	1	1	0
2020-06-14	1	1	0
2020-06-21	1	1	0
2020-06-28	1	1	0
2020-07-05	1	1	0
2020-07-12	1	1	0
2020-07-19	1	1	1
2020-07-26	1	1	0
2020-08-02	1	1	1
2020-08-09	1	1	0
2020-08-16	1	0	0
2020-08-23	1	0	0

Table S4 is not appropriate for this report and so has been omitted.

Table S5 Raw data for figure five showing when lineages started per day, divided by singletons and non-singletons

Day	Number of singleton starts	Number of non-singleton starts	Total
2020-03-01	0	1	1
2020-03-02	0	3	3
2020-03-03	0	1	1
2020-03-05	0	1	1
2020-03-06	0	1	1
2020-03-07	0	1	1
2020-03-08	1	1	2
2020-03-12	1	0	1
2020-03-13	0	1	1
2020-03-15	0	1	1
2020-03-16	0	1	1
2020-03-17	0	1	1
2020-03-18	0	5	5
2020-03-19	1	1	2
2020-03-20	2	3	5
2020-03-21	2	5	7
2020-03-22	2	3	5
2020-03-23	3	3	6
2020-03-24	8	5	13
2020-03-25	10	10	20
2020-03-26	8	4	12
2020-03-27	2	2	4
2020-03-28	1	2	3
2020-03-29	4	2	6
2020-03-30	0	2	2
2020-03-31	2	2	4
2020-04-01	4	1	5
2020-04-02	5	2	7
2020-04-03	3	0	3
2020-04-04	3	1	4

Day	Number of singleton starts	Number of non-singleton starts	Total
2020-04-05	5	1	6
2020-04-06	1	1	2
2020-04-07	2	1	3
2020-04-08	1	0	1
2020-04-09	3	0	3
2020-04-10	1	1	2
2020-04-11	2	1	3
2020-04-13	1	0	1
2020-04-14	1	0	1
2020-04-15	3	1	4
2020-04-16	1	0	1
2020-04-17	2	0	2
2020-04-18	1	0	1
2020-04-19	1	1	2
2020-04-20	1	0	1
2020-04-21	1	0	1
2020-04-22	1	1	2
2020-04-23	1	0	1
2020-04-27	1	0	1
2020-04-28	1	0	1
2020-05-04	0	1	1
2020-05-05	1	0	1
2020-05-10	1	0	1
2020-05-19	0	1	1
2020-05-25	0	2	2
2020-05-26	1	1	2
2020-05-27	2	0	2
2020-06-08	1	0	1
2020-06-25	1	0	1
2020-07-03	1	1	2
2020-07-18	1	0	1
2020-08-22	1	0	1

Table S6 Raw data for figure six showing the number of sequences taken over time.

Day	England
2020-03-01	1
2020-03-02	4
2020-03-03	2
2020-03-05	1
2020-03-06	1
2020-03-07	1
2020-03-08	2
2020-03-09	2
2020-03-12	4
2020-03-13	1
2020-03-15	3
2020-03-16	4
2020-03-17	4
2020-03-18	16
2020-03-19	9
2020-03-20	16
2020-03-21	27
2020-03-22	24
2020-03-23	33
2020-03-24	57
2020-03-25	78
2020-03-26	63

Day	England
2020-03-27	36
2020-03-28	31
2020-03-29	30
2020-03-30	56
2020-03-31	32
2020-04-01	56
2020-04-02	75
2020-04-03	44
2020-04-04	31
2020-04-05	18
2020-04-06	34
2020-04-07	19
2020-04-08	28
2020-04-09	39
2020-04-10	43
2020-04-11	43
2020-04-12	23
2020-04-13	17
2020-04-14	27
2020-04-15	29
2020-04-16	25
2020-04-17	37
2020-04-18	36
2020-04-19	28
2020-04-20	30
2020-04-21	25
2020-04-22	34
2020-04-23	21
2020-04-24	15
2020-04-25	10
2020-04-26	19
2020-04-27	36
2020-04-28	18
2020-04-29	33
2020-04-30	16
2020-05-01	18
2020-05-02	15
2020-05-03	12
2020-05-04	22
2020-05-05	16
2020-05-06	13
2020-05-07	15
2020-05-08	15
2020-05-09	5
2020-05-10	18
2020-05-11	19
2020-05-12	11
2020-05-13	16
2020-05-14	10
2020-05-15	12
2020-05-16	11
2020-05-17	4
2020-05-18	13
2020-05-19	15
2020-05-20	7
2020-05-21	6
2020-05-22	10
2020-05-23	2
2020-05-24	3

Day	England
2020-05-25	9
2020-05-26	14
2020-05-27	33
2020-05-28	17
2020-05-29	9
2020-05-30	11
2020-06-01	4
2020-06-02	5
2020-06-03	6
2020-06-04	3
2020-06-05	3
2020-06-06	2
2020-06-07	4
2020-06-08	4
2020-06-09	5
2020-06-10	6
2020-06-11	3
2020-06-12	6
2020-06-13	6
2020-06-14	6
2020-06-15	9
2020-06-16	3
2020-06-17	1
2020-06-18	3
2020-06-19	4
2020-06-20	1
2020-06-21	2
2020-06-22	3
2020-06-23	1
2020-06-24	4
2020-06-25	2
2020-06-26	1
2020-06-27	1
2020-06-28	1
2020-06-29	2
2020-07-01	1
2020-07-03	8
2020-07-04	1
2020-07-06	3
2020-07-07	1
2020-07-09	1
2020-07-10	7
2020-07-11	1
2020-07-12	2
2020-07-14	4
2020-07-17	2
2020-07-18	1
2020-07-19	1
2020-07-20	1
2020-07-21	8
2020-07-23	2
2020-07-24	1
2020-07-25	1
2020-07-26	1
2020-07-30	5
2020-08-02	1
2020-08-03	3
2020-08-04	2
2020-08-06	2

Day	England
2020-08-08	1
2020-08-10	3
2020-08-11	5
2020-08-12	1
2020-08-13	1
2020-08-15	4
2020-08-17	1
2020-08-19	1
2020-08-22	1
2020-08-24	1
2020-08-25	1

Table S7 Raw data for the figure seven with the number of sequences assigned to each admin2 region.

Admin2	Country	Number of sequences	Sequence group
DERBYSHIRE	England	27	10-100
EAST RIDING OF YORKSHIRE	England	8	1-10
SOUTH YORKSHIRE	England	1970	1000-2000