UK lineages summary report

This report gives summaries of UK specific lineages sequenced by PHEC for week 2020-05-29. There are time lags due to batching, curation and analysis, the most recently sampled sequence is 2020-05-05. The analysis (eg time since last sample) is therefore undertaken from this date. 3847 sequences in the UK from the sequencing centre PHEC 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 6130 and the maximum is 9084

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

1944 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 47 that remain: 31 are pending extinction, ie last seen three weeks ago. 2 lineages have gone quiet, ie haven't been seen this week. 11 lineages have reactivated. 3 lineages have been continuously circulating.

The following table contains information about the ten largest lineages lineages and the number of sequences the dataset. Information about other lienages 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 expect 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 2020-05-19

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	Total sequences	Global lineage	Time since last sample (days)	Activity score
UK701	124 (100.0%)	Feb-03, Apr-30	124	B.1, B.1.p11	5	0.1415
UK5	97 (100.0%)	Mar-03, May-05	97	B.1.1.1	0	active today
UK9	82 (100.0%)	Mar-09, Apr-17	82	B.1.13	18	0.0267
UK107	68 (100.0%)	Mar-15, Apr-21	68	B.2, B.2.1, B.2.5	14	0.0394
UK2464	49 (100.0%)	Mar-09, May-04	49	B.1.p11	1	1.1667
UK77	48 (100.0%)	Mar-11, May-05	48	B.2, B.2.4	0	active today

Lineage name	England	Date range	Total sequences	Global lineage	Time since last sample (days)	Activity score
UK63	39 (100.0%)	Mar-18, May-04	39	B.1.1	1	1.2368
UK19	35 (100.0%)	Mar-09, Apr-21	35	B.1	14	0.0903
UK116	28 (100.0%)	Feb-25, Apr-01	28	B.2.1	34	0.0392
UK94	28 (100.0%)	Mar-12, Apr-19	28	B.2, B.2.1	16	0.088

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.

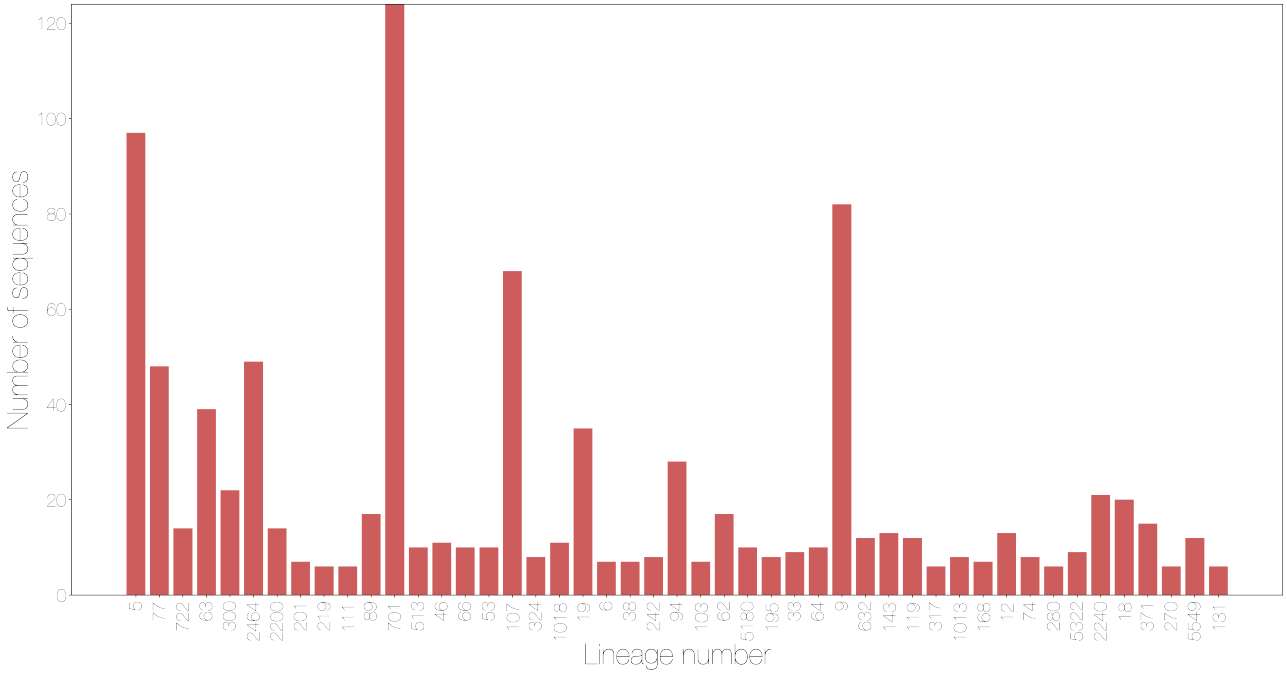


Figure 1: Number of sequences sampled in a lineage by country

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.

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.

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.

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

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.

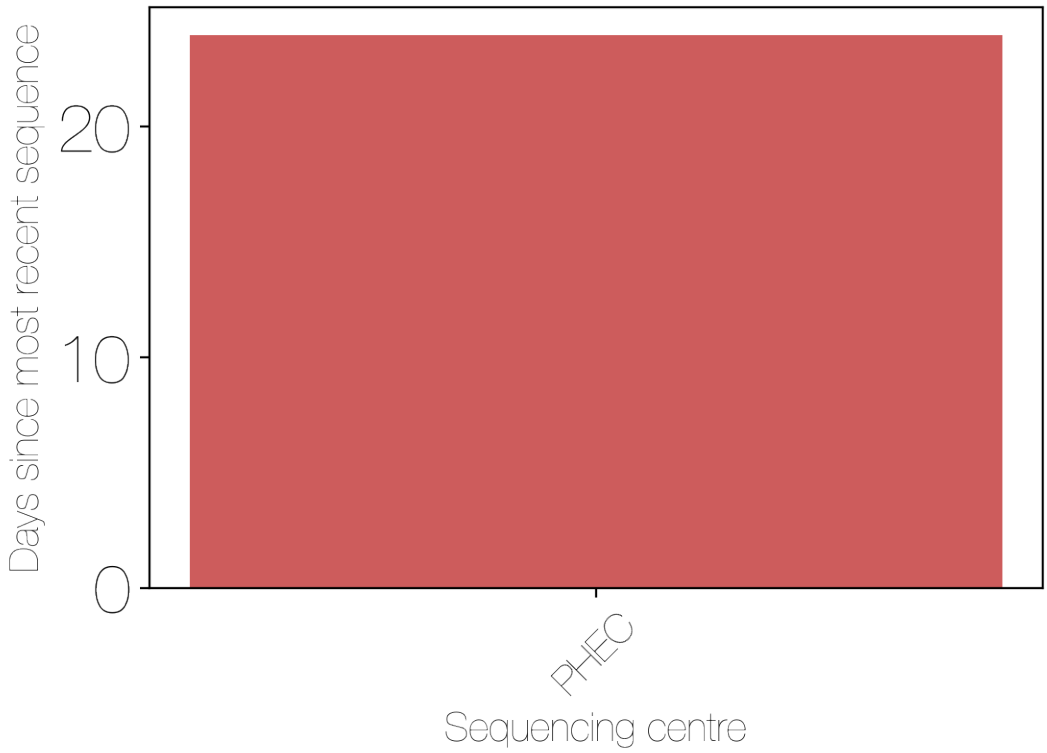


Figure 2: Lag since the most recent sequence from each sequencing centre to most current date

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 950 sequences without enough geographical information to map from this centre.

———————————————————————————FileNotFoundError Traceback (most recent call last) in 5 input_geojsons = [uk_json, channels, NI_json] 6 —-> 7 map_output = map.make_map(input_geojsons, adm2_cleaning_file, metadata_file, output_directory, week, sequencing_centre, country) 8 9 if type(map_output) != bool: ~/anaconda3/envs/report/lib/python3.7/site-packages/UK_full_report/utils/mapping.py in make_map(input_geojs adm2_cleaning_file, metadata_file, overall_output_dir, week, sequencing_centre, country) 540 sort_missing_sequences(mis missing_sequences, sequencing_centre, country) 541 –> 542 new_unclean_locs = find_new_locs_cleaning(metadata_file, mapping_dictionary, all_uk, output_dir, sequencing_centre) 543 544 return new_unclean_locs, cleaned ~/anaconda3/envs/report/lib/python3.7/site-packages/UK_full_report/utils/mapping.py in find_new_locs_cleaning(metada mapping_dictionary, all_uk, output_dir, sequencing_centre) 426 427 new_unclean_locs = False –> 428 fw = open(output_dir + “unclean_locations.csv”, ‘w’) 429 430 for i in all_uk[“NAME_2”]: FileNotFoundError: [Errno 2] No such file or directory: ‘UK_full_report/regional_reports/results/results_PHEC/summary_files/unclean_locations.csv’

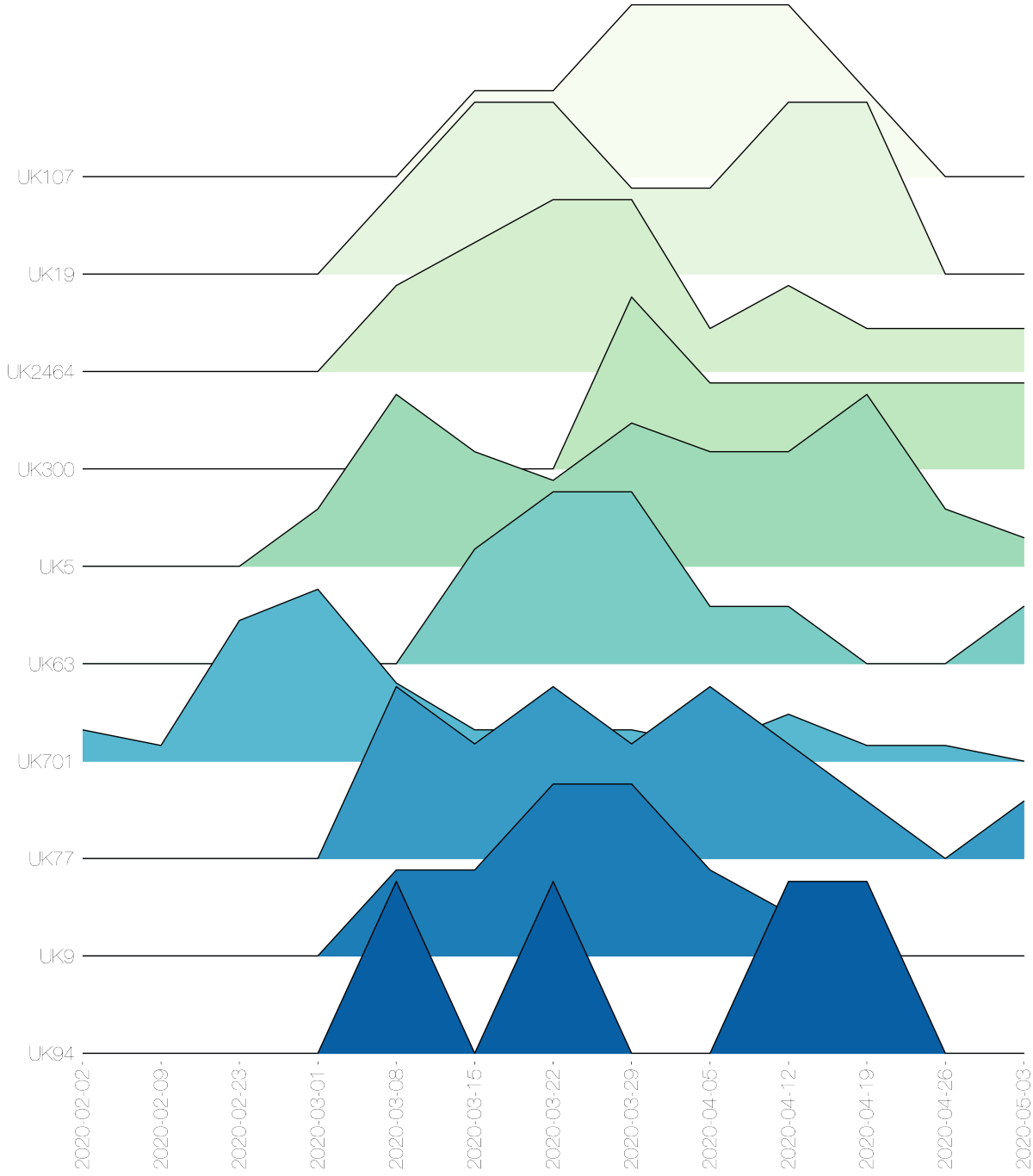


Figure 3: Lineages by number of adm2 regions present by epiweek

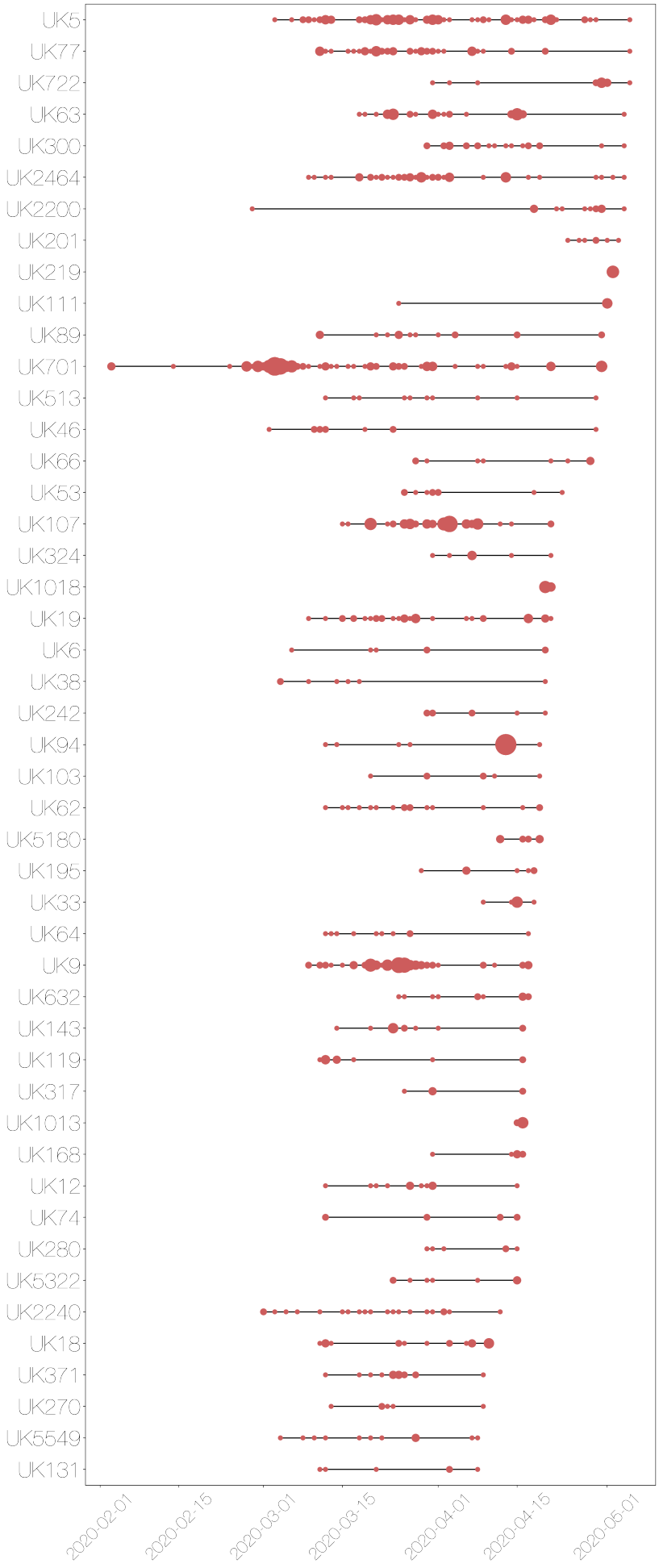


Figure 4: Timeline of lineages, sized by number of sequences from each country.

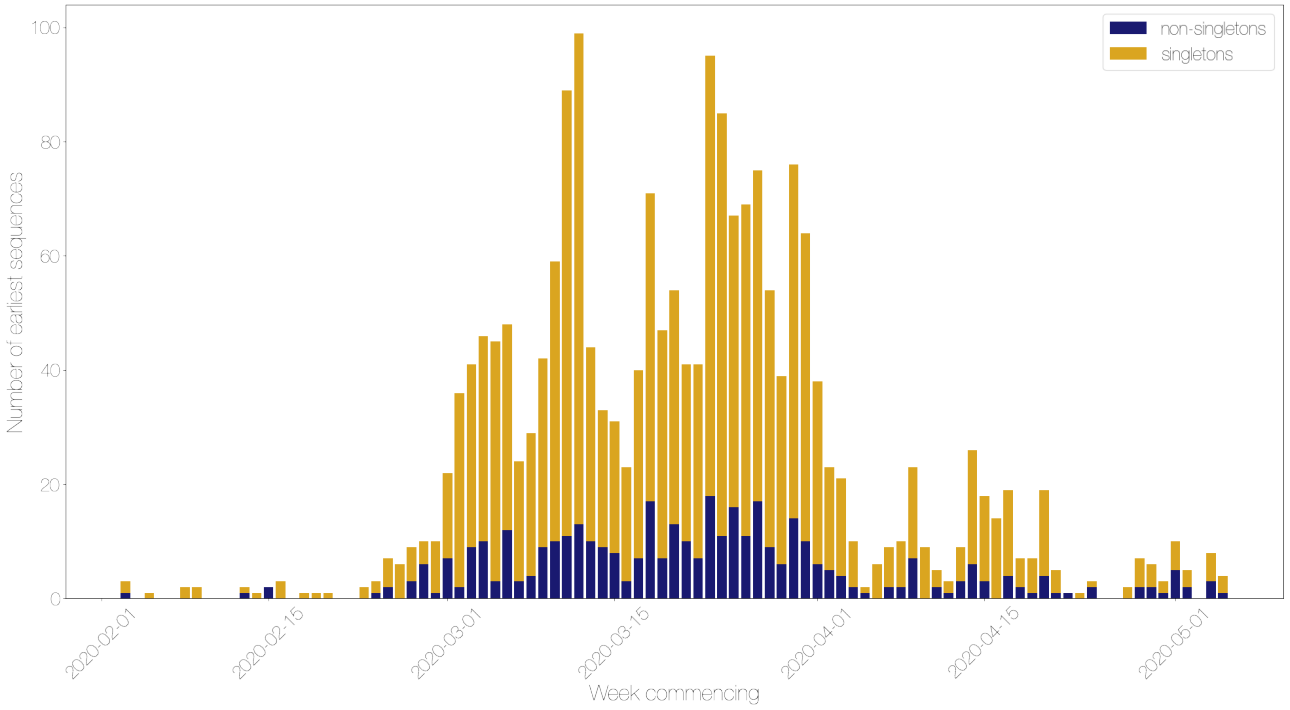


Figure 5: Lineage starts per week, split by singletons and non-singletons

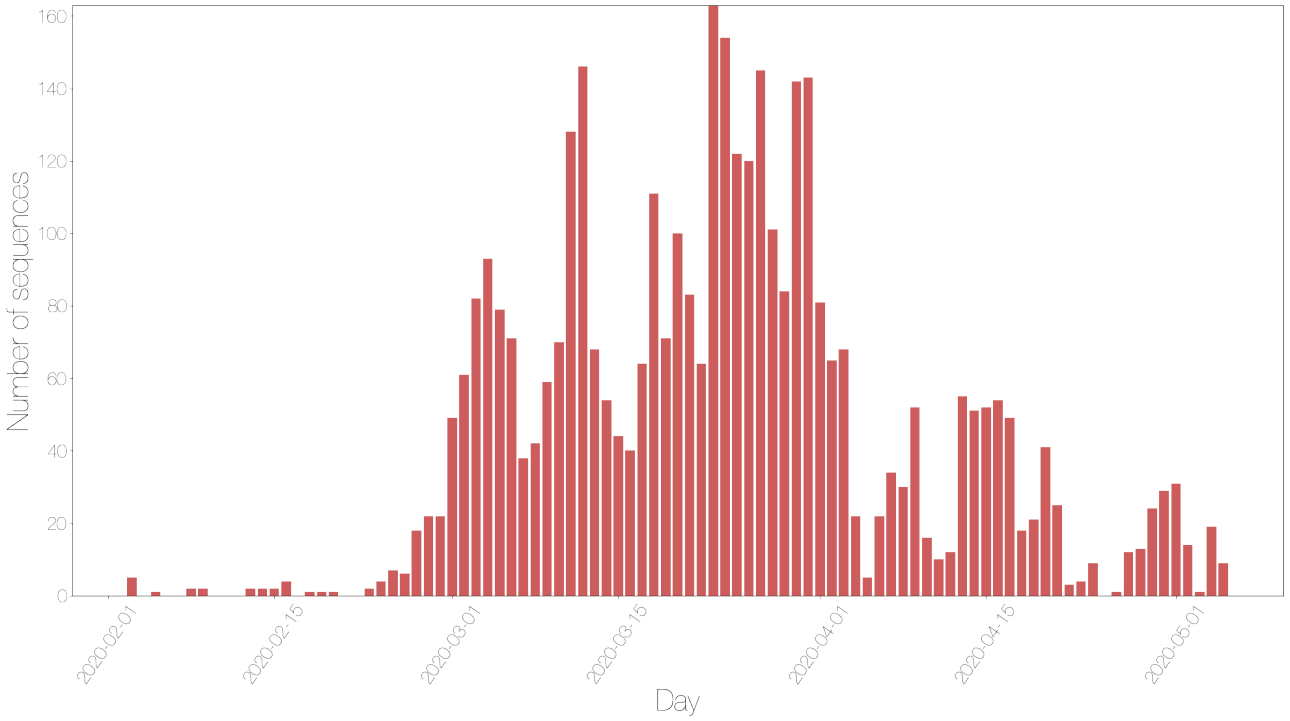
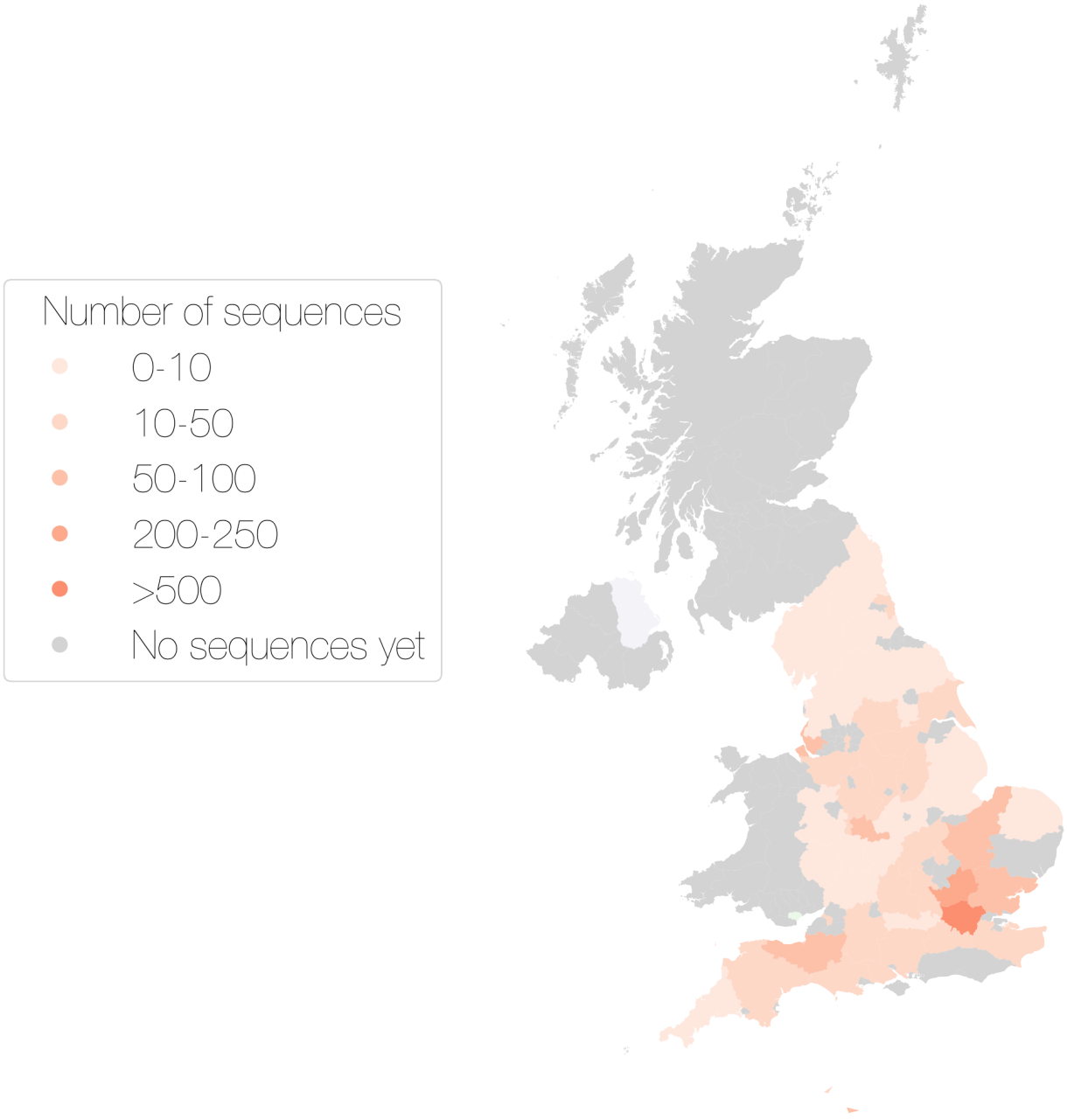


Figure 6: Sequences taken on each day by country



———————————————————————————NameError Traceback (most recent call last) in —-> 1 if not no_seqs: 2 if new_uncleans: 3 print("There are some sequences with locations that are not matched to real Admin2 regions, some manual curation required.") NameError: name 'no_seqs' is not defined

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	Total sequences	Global lineage	Time since last sample (days)	Activity score
UK701	124 (100.0%)	Feb-03, Apr-30	124	B.1, B.1.p11	5	0.1415
UK5	97 (100.0%)	Mar-03, May-05	97	B.1.1.1	0	active today
UK9	82 (100.0%)	Mar-09, Apr-17	82	B.1.13	18	0.0267
UK107	68 (100.0%)	Mar-15, Apr-21	68	B.2, B.2.1, B.2.5	14	0.0394
UK2464	49 (100.0%)	Mar-09, May-04	49	B.1.p11	1	1.1667
UK77	48 (100.0%)	Mar-11, May-05	48	B.2, B.2.4	0	active today
UK63	39 (100.0%)	Mar-18, May-04	39	B.1.1	1	1.2368
UK19	35 (100.0%)	Mar-09, Apr-21	35	B.1	14	0.0903
UK116	28 (100.0%)	Feb-25, Apr-01	28	B.2.1	34	0.0392
UK94	28 (100.0%)	Mar-12, Apr-19	28	B.2, B.2.1	16	0.088
UK4	26 (100.0%)	Feb-28, Mar-31	26	B	35	0.0366
UK3	24 (100.0%)	Feb-24, Apr-05	24	B.1	30	0.0594
UK300	22 (100.0%)	Mar-30, May-04	22	B.1.1	1	1.6667
UK2240	21 (100.0%)	Mar-01, Apr-12	21	B.1	23	0.0913
UK18	20 (100.0%)	Mar-11, Apr-10	20	B.1.1.7	25	0.0632
UK41	19 (100.0%)	Mar-01, Mar-26	19	B.1	40	0.0347
UK62	17 (100.0%)	Mar-12, Apr-19	17	B.3	16	0.1484
UK89	17 (100.0%)	Mar-11, Apr-30	17	B.1.1.9	5	0.625
UK37	17 (100.0%)	Mar-18, Apr-02	17	B.1, B.1.30	33	0.0284
UK241	16 (100.0%)	Mar-25, Apr-07	16	B.1.5.3	28	0.031
UK1845	16 (100.0%)	Mar-01, Mar-31	16	B	35	0.0571

Lineage name	England	Date range	Total sequences	Global lineage	Time since last sample (days)	Activity score
UK371	15 (100.0%)	Mar-12, Apr-09	15	B.1.1	26	0.0769
UK274	15 (100.0%)	Mar-06, Apr-02	15	B, B.3	33	0.0584
UK117	15 (100.0%)	Feb-28, Mar-22	15	B.2.1	44	0.0373
UK2200	14 (100.0%)	Feb-28, May-04	14	B.1.5.6, B.1.5	1	5.0769
UK112	14 (100.0%)	Mar-15, Mar-31	14	B.1.1	35	0.0352
UK722	14 (100.0%)	Mar-31, May-05	14	B.1.1	0	active today
UK403	14 (100.0%)	Mar-23, Mar-31	14	B.1.1	35	0.0176
UK143	13 (100.0%)	Mar-14, Apr-16	13	B.2.1	19	0.1447
UK12	13 (100.0%)	Mar-12, Apr-15	13	B.1.p11	20	0.1417
UK378	13 (100.0%)	Feb-15, Mar-05	13	B.1.1	61	0.026
UK34	13 (100.0%)	Feb-15, Apr-02	13	B.4	33	0.1187
UK347	12 (100.0%)	Mar-13, Apr-02	12	B.1	33	0.0551
UK119	12 (100.0%)	Mar-11, Apr-16	12	B.2.5	19	0.1722
UK632	12 (100.0%)	Mar-25, Apr-17	12	B.1.1	18	0.1162
UK5549	12 (100.0%)	Mar-04, Apr-08	12	B.2.2	27	0.1178
UK694	12 (100.0%)	Mar-06, Mar-14	12	B	52	0.014
UK1018	11 (100.0%)	Apr-20, Apr-21	11	B.1.1	14	0.0071
UK46	11 (100.0%)	Mar-02, Apr-29	11	B.2.1	6	0.9667
UK291	10 (100.0%)	Mar-13, Apr-03	10	B.2.1	32	0.0729
UK64	10 (100.0%)	Mar-12, Apr-17	10	B.1	18	0.2222
UK513	10 (100.0%)	Mar-12, Apr-29	10	B.1.p11	6	0.8889
UK428	10 (100.0%)	Mar-20, Apr-01	10	B.2, B.2.1	34	0.0392
UK66	10 (100.0%)	Mar-28, Apr-28	10	B.1.1.8	7	0.4921
UK5180	10 (100.0%)	Apr-12, Apr-19	10	B.1.1.7	16	0.0486

Lineage name	England	Date range	Total sequences	Global lineage	Time since last sample (days)	Activity score
UK604	10 (100.0%)	Mar-09, Mar-12	10	B.1.1	54	0.0062
UK5715	10 (100.0%)	Feb-13, Mar-07	10	B.2	59	0.0433
UK494	10 (100.0%)	Mar-26, Apr-01	10	B.1.p11	34	0.0196
UK53	10 (100.0%)	Mar-26, Apr-23	10	B.1.1.4	12	0.2593
UK33	9 (100.0%)	Apr-09, Apr-18	9	B.1.1	17	0.0662
UK5322	9 (100.0%)	Mar-24, Apr-15	9	B.1.1	20	0.1375
UK687	9 (100.0%)	Feb-28, Mar-08	9	B.2, B.2.1	58	0.0194
UK190	9 (100.0%)	Mar-01, Mar-30	9	B.1	36	0.1007
UK739	8 (100.0%)	Mar-01, Mar-08	8	B.4	58	0.0172
UK74	8 (100.0%)	Mar-12, Apr-15	8	B.1	20	0.2429
UK1013	8 (100.0%)	Apr-15, Apr-16	8	B.1.1	19	0.0075
UK195	8 (100.0%)	Mar-29, Apr-18	8	B.1.1	17	0.1681
UK11	8 (100.0%)	Mar-06, Mar-25	8	B.1	41	0.0662
UK155	8 (100.0%)	Feb-27, Mar-24	8	B.1	42	0.0884
UK242	8 (100.0%)	Mar-30, Apr-20	8	B.1.5	15	0.2
UK324	8 (100.0%)	Mar-31, Apr-21	8	B.1.1	14	0.2143
UK756	8 (100.0%)	Feb-27, Mar-05	8	B.1.1	61	0.0164
UK788	8 (100.0%)	Feb-28, Mar-05	8	B.4	61	0.0141
UK168	7 (100.0%)	Mar-31, Apr-16	7	B.2.1	19	0.1404
UK201	7 (100.0%)	Apr-24, May-03	7	B.1	2	0.75
UK6	7 (100.0%)	Mar-06, Apr-20	7	B.1	15	0.5
UK8	7 (100.0%)	Mar-03, Mar-12	7	B	54	0.0278
UK733	7 (100.0%)	Mar-10, Mar-18	7	B.2.1	48	0.0278
UK22	7 (100.0%)	Mar-02, Mar-18	7	B	48	0.0556

Lineage name	England	Date range	Total sequences	Global lineage	Time since last sample (days)	Activity score
UK38	7 (100.0%)	Mar-04, Apr-20	7	B.2.1	15	0.5222
UK103	7 (100.0%)	Mar-20, Apr-19	7	B.1.1	16	0.3125
UK223	6 (100.0%)	Mar-10, Mar-27	6	B.2.1	39	0.0872
UK171	6 (100.0%)	Mar-13, Mar-27	6	B.2, B.2.1	39	0.0718
UK111	6 (100.0%)	Mar-25, May-01	6	B.1.1	4	1.85
UK335	6 (100.0%)	Mar-25, Mar-31	6	B.2.1	35	0.0343
UK317	6 (100.0%)	Mar-26, Apr-16	6	B.3	19	0.2211
UK857	6 (100.0%)	Mar-24, Mar-29	6	B.2.1	37	0.027
UK131	6 (100.0%)	Mar-11, Apr-08	6	B.15	27	0.2074
UK799	6 (100.0%)	Mar-01, Mar-07	6	B.1	59	0.0203
UK280	6 (100.0%)	Mar-30, Apr-15	6	B.1.1	20	0.16
UK219	6 (100.0%)	May-02, May-02	6	B.1.1	3	0.0
UK178	6 (100.0%)	Mar-14, Apr-04	6	B.1.1	31	0.1355
UK654	6 (100.0%)	Feb-27, Mar-08	6	B.2.5	58	0.0345
UK270	6 (100.0%)	Mar-13, Apr-09	6	B.3	26	0.2077

Table S2 Raw data for figure three showing the number of admin2 regions a lineage is present in over time

Week commencing	UK701	UK5	UK9	UK107	UK2464	UK77	UK63	UK19	UK94	UK300
2020-02-02	2	0	0	0	0	0	0	0	0	0
2020-02-09	1	0	0	0	0	0	0	0	0	0
2020-02-23	9	0	0	0	0	0	0	0	0	0
2020-03-01	11	2	0	0	0	0	0	0	0	0
2020-03-08	5	6	2	0	2	3	0	1	1	0
2020-03-15	2	4	2	1	3	2	2	2	0	0
2020-03-22	2	3	4	1	4	3	3	2	1	0
2020-03-29	2	5	4	2	4	2	3	1	0	2
2020-04-05	1	4	2	2	1	3	1	1	0	1
2020-04-12	3	4	1	2	2	2	1	2	1	1
2020-04-19	1	6	0	1	1	1	0	2	1	1
2020-04-26	1	2	0	0	1	0	0	0	0	1
2020-05-03	0	1	0	0	1	1	1	0	0	1

Table S3 is not appropriate for this report and so has been omitted.

Table S4 Raw data for figure six showing when lineages started per day, divided by singletons and non-singletons

Day	Number of singleton starts	Number of non-singleton starts	Total
2020-02-03	2	1	3
2020-02-05	1	0	1
2020-02-08	2	0	2
2020-02-09	2	0	2
2020-02-13	1	1	2
2020-02-14	1	0	1
2020-02-15	0	2	2
2020-02-16	3	0	3
2020-02-18	1	0	1
2020-02-19	1	0	1
2020-02-20	1	0	1
2020-02-23	2	0	2
2020-02-24	2	1	3
2020-02-25	5	2	7
2020-02-26	6	0	6
2020-02-27	6	3	9
2020-02-28	4	6	10
2020-02-29	9	1	10
2020-03-01	15	7	22
2020-03-02	34	2	36
2020-03-03	32	9	41
2020-03-04	36	10	46
2020-03-05	42	3	45
2020-03-06	36	12	48
2020-03-07	21	3	24
2020-03-08	25	4	29
2020-03-09	33	9	42
2020-03-10	49	10	59
2020-03-11	78	11	89
2020-03-12	86	13	99
2020-03-13	34	10	44
2020-03-14	24	9	33
2020-03-15	23	8	31
2020-03-16	20	3	23
2020-03-17	33	7	40
2020-03-18	54	17	71
2020-03-19	40	7	47
2020-03-20	41	13	54
2020-03-21	31	10	41
2020-03-22	34	7	41
2020-03-23	77	18	95
2020-03-24	74	11	85
2020-03-25	51	16	67
2020-03-26	58	11	69
2020-03-27	58	17	75
2020-03-28	45	9	54

Day	Number of singleton starts	Number of non-singleton starts	Total
2020-03-29	33	6	39
2020-03-30	62	14	76
2020-03-31	54	10	64
2020-04-01	32	6	38
2020-04-02	18	5	23
2020-04-03	17	4	21
2020-04-04	8	2	10
2020-04-05	1	1	2
2020-04-06	6	0	6
2020-04-07	7	2	9
2020-04-08	8	2	10
2020-04-09	16	7	23
2020-04-10	9	0	9
2020-04-11	3	2	5
2020-04-12	2	1	3
2020-04-13	6	3	9
2020-04-14	20	6	26
2020-04-15	15	3	18
2020-04-16	14	0	14
2020-04-17	15	4	19
2020-04-18	5	2	7
2020-04-19	6	1	7
2020-04-20	15	4	19
2020-04-21	4	1	5
2020-04-22	0	1	1
2020-04-23	1	0	1
2020-04-24	1	2	3
2020-04-27	2	0	2
2020-04-28	5	2	7
2020-04-29	4	2	6
2020-04-30	2	1	3
2020-05-01	5	5	10
2020-05-02	3	2	5
2020-05-04	5	3	8
2020-05-05	3	1	4

Table S5 Raw data for figure seven showing the number of sequences taken over time.

Day	England
2020-02-03	5
2020-02-05	1
2020-02-08	2
2020-02-09	2
2020-02-13	2
2020-02-14	2
2020-02-15	2
2020-02-16	4
2020-02-18	1
2020-02-19	1
2020-02-20	1
2020-02-23	2
2020-02-24	4
2020-02-25	7
2020-02-26	6
2020-02-27	18
2020-02-28	22
2020-02-29	22
2020-03-01	49
2020-03-02	61
2020-03-03	82
2020-03-04	93
2020-03-05	79
2020-03-06	71
2020-03-07	38
2020-03-08	42
2020-03-09	59
2020-03-10	70
2020-03-11	128
2020-03-12	146
2020-03-13	68
2020-03-14	54
2020-03-15	44
2020-03-16	40
2020-03-17	64
2020-03-18	111
2020-03-19	71
2020-03-20	100
2020-03-21	83
2020-03-22	64
2020-03-23	163
2020-03-24	154
2020-03-25	122
2020-03-26	120
2020-03-27	145
2020-03-28	101
2020-03-29	84

Day	England
2020-03-30	142
2020-03-31	143
2020-04-01	81
2020-04-02	65
2020-04-03	68
2020-04-04	22
2020-04-05	5
2020-04-06	22
2020-04-07	34
2020-04-08	30
2020-04-09	52
2020-04-10	16
2020-04-11	10
2020-04-12	12
2020-04-13	55
2020-04-14	51
2020-04-15	52
2020-04-16	54
2020-04-17	49
2020-04-18	18
2020-04-19	21
2020-04-20	41
2020-04-21	25
2020-04-22	3
2020-04-23	4
2020-04-24	9
2020-04-26	1
2020-04-27	12
2020-04-28	13
2020-04-29	24
2020-04-30	29
2020-05-01	31
2020-05-02	14
2020-05-03	1
2020-05-04	19
2020-05-05	9

Table S6 Raw data for the map with the number of sequences assigned to each admin2 region.

———————————————————————————NameError Traceback (most recent call last) in —-> 1 if not no_seqs: 2 print(mapping_data.to_markdown()) NameError: name ‘no_seqs’ is not defined

```
---------------------------------------------------------------------------FileNotFoundError
Traceback (most recent call last)<ipython-input-1-c2b516fe2325> in
<module>
----> 1 writing.write_summary_files(summary_output, dataframe,
omitted, week, intro_alls, timeline_df)
~/anaconda3/envs/report/lib/python3.7/site-
packages/UK_full_report/utils/writing_summary_files.py in
write_summary_files(output_dir, dataframe, omitted, week, intro_alls,
timeline_data)
     55 def write_summary_files(output_dir, dataframe, omitted, week,
intro_alls, timeline_data):
     56
---> 57     write_summary_table(dataframe, output_dir)
     58     write_omitteds(omitted, output_dir)
     59     write_singletons(intro_alls, output_dir)
~/anaconda3/envs/report/lib/python3.7/site-
packages/UK_full_report/utils/writing_summary_files.py in
write_summary_table(dataframe, output_dir)
      4 def write_summary_table(dataframe, output_dir):
      5
----> 6     dataframe.to_csv(output_dir + "/lineage_summary.tsv",
sep="\t")
      7
      8 def write_all_lins(intro_alls, output_dir):
~/anaconda3/envs/report/lib/python3.7/site-
packages/pandas/core/generic.py in to_csv(self, path_or_buf, sep,
na_rep, float_format, columns, header, index, index_label, mode,
encoding, compression, quoting, quotechar, line_terminator, chunksize,
date_format, doublequote, escapechar, decimal)
   3202             decimal=decimal,
   3203         )
-> 3204         formatter.save()
   3205
   3206         if path_or_buf is None:
~/anaconda3/envs/report/lib/python3.7/site-
packages/pandas/io/formats/csvs.py in save(self)
    186                 self.mode,
    187                 encoding=self.encoding,
--> 188                 compression=dict(self.compression_args,
method=self.compression),
    189             )
    190             close = True
~/anaconda3/envs/report/lib/python3.7/site-
packages/pandas/io/common.py in get_handle(path_or_buf, mode,
encoding, compression, memory_map, is_text)
    426         if encoding:
    427             # Encoding
--> 428             f = open(path_or_buf, mode, encoding=encoding,
newline="")
    429         elif is_text:
```

```
    430             # No explicit encoding
FileNotFoundError: [Errno 2] No such file or directory:
'UK_full_report/regional_reports/results/results_PHEC/summary_files/lineage_summary.tsv'
```