COVID-19 UK introductions

Transmission lineage summary

Louis du Plessis

Last modified: 07 Dec 2020

Contents

1 Summary

This notebook plots summary statistics and figures of the UK transmission lineages extracted from the BEAST DTA analyses.

1.1 Input

- Metadata table (in `inputpath`).
- Cluster statistics for MCC trees and across posterior trees as produced (in `outputpath`):
  - `clusters_DTA.csv`
  - `clusterSamples_DTA.csv`
  - `clusters_DTA_MCC_0.5.csv`
  - `clusterSamples_DTA_MCC_0.5.csv`
- Combined `.log` file from DTA analysis (in `logpath`).

1.2 Output

- Lineage summary figures and tables.
- Distributions of transmission lineage sizes in the MCC trees and across posterior trees (as `.csv` files).

2 Transmission lineage statistics

2.1 Input data

- DTA transmission lineages and singletons, on the dataset from 26 June, n = 50887 sequences (n = 26181, 51% from the UK).
- Oldest sequence: 2019-12-24
- Newest sequence: 2020-06-22

2.2 Summary statistics (2000 posterior trees)

- TMRCAs were estimated across 2000 posterior trees using BEAST with a fixed clock-rate and DTA was used to identify transmission lineages and singletons.
- Dataset contains 1219 [1143,1286] UK transmission lineages (2 or more sequences), comprising 24483 [24398,24570] sequences from the UK, as well as a further 1698 [1611,1783] singletons.
- Mean and SD of the median TMRCA distributions across 2000 posterior trees: 2020-03-22 ± 14.888 days (singletons excluded).
- Median and interquartile range of TMRCA distribution across 2000 posterior trees: 2020-03-21 [2020-03-13, 2020-03-29] (singletons excluded).
- 868 [807,927] small lineages (<10 sequences), making up 71.21% [69.31,72.94] of all transmission lineages.

2.3 Summary statistics (MCC tree)

- Built MCC tree from 2000 posterior trees and used a threshold of 0.5 posterior probability to identify internal nodes in the UK (and identify transmission lineages).
- Dataset contains 1179 UK transmission lineages (2 or more sequences), comprising 24531 sequences from the UK, as well as a further 1650 singletons.
- Mean and SD of the TMRCA distribution: 2020-03-22 ± 14.67 days (singletons excluded).

- Median and interquartile range of TMRCA distribution: 2020-03-21 [2020-03-14, 2020-03-29] (singletons excluded).
- 854 small lineages (<10 sequences), making up 72.43% of all transmission lineages.

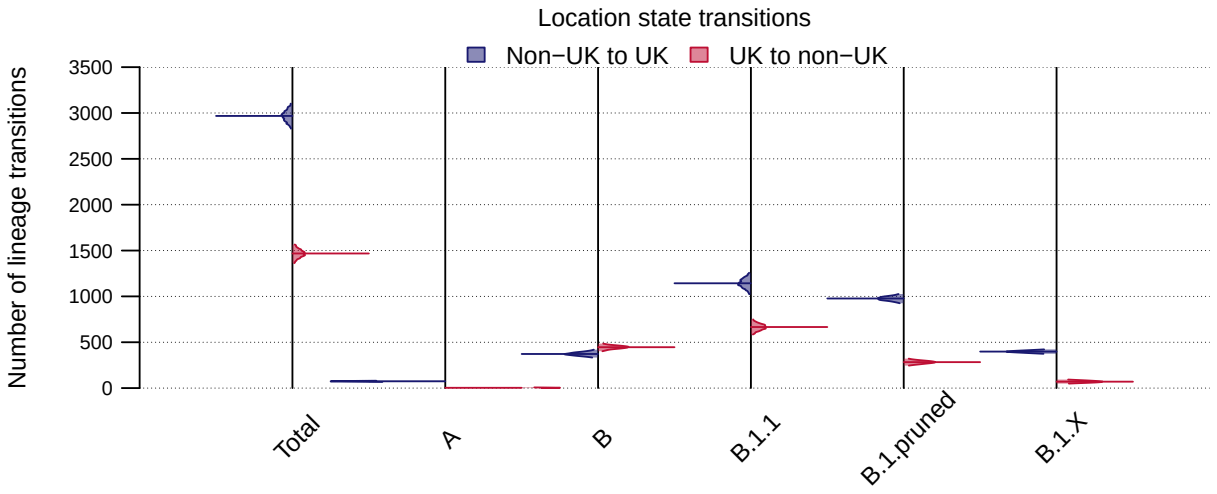

Figure 1: Number of location state transitions between the binary phylogenetic traits UK/non-UK detected by the robust counting approach implemented in BEAST 1.10. Non-UK to UK=blue, UK to non-UK=red. Posterior distributions are truncated at their 95% HPD interval limits and the horizontal lines indicate median estimates.

```
## 548.81 sec elapsed
```

Table 1: The number of location state transitions (non-UK to UK and vice-versa) taken across the set of 2000 posterior trees, as well as the total number of transmission lineages and singletons inferred across the set of 2000 posterior trees and the MCC trees. Numbers are given for the whole dataset and for each individual subtree.

	Non-UK to UK state transitions (median and 95% HPD)		UK to non-UK state transitions (median and 95% HPD)		Transmission lineages and singletons (median and 95% HPD)		Transmission lineages and singletons in MCC tree
Total	2968	[2829-3103]	1468	[1362-1566]	2918	[2773-3048]	2829
A	74	[67-80]	3	[0-7]	74	[67-80]	69
B	372	[333-419]	446	[402-485]	365	[326-409]	360
B.1.1	1143	[1023-1258]	666	[584-749]	1115	[992-1230]	1074
B.1.pruned	977	[925-1026]	283	[245-321]	964	[914-1015]	943
B.1.X	398	[374-422]	70	[49-93]	396	[370-419]	383

- The 20% biggest transmission lineages contain 76.04 [73.08, 78.86] of all UK genomes (across all 2000

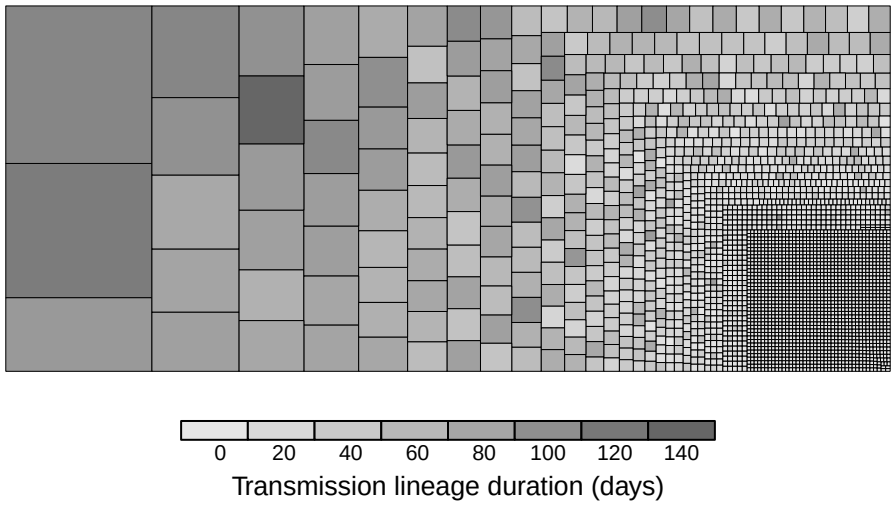

Figure 2: Partition of 26,181 UK genomes into UK transmission lineages and singletons, coloured by duration of lineage detection (time between the lineage's oldest and most recent genomes). (Transmission lineages from the MCC trees).

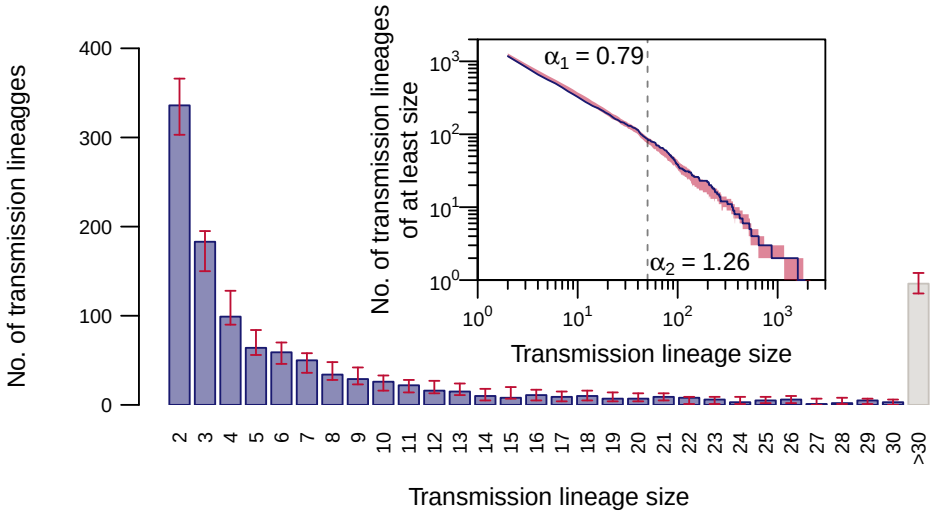

Figure 3: Distribution of UK transmission lineage sizes (MCC trees). Blue bars show the number of transmission lineages of each size (red bars=95% HPD of these sizes across the posterior tree distribution). Inset: the corresponding complementary cumulative frequency distribution of lineage size (blue line), on double logarithmic axes (red shading=95% HPD of this distribution across the posterior tree distribution). Values either side of vertical dashed line show coefficients of power-law distributions (P[X >= x] ~ xˆalpha)) fitted to lineages containing <=50 (alpha1) and >50 (alpha2) virus genomes, respectively.

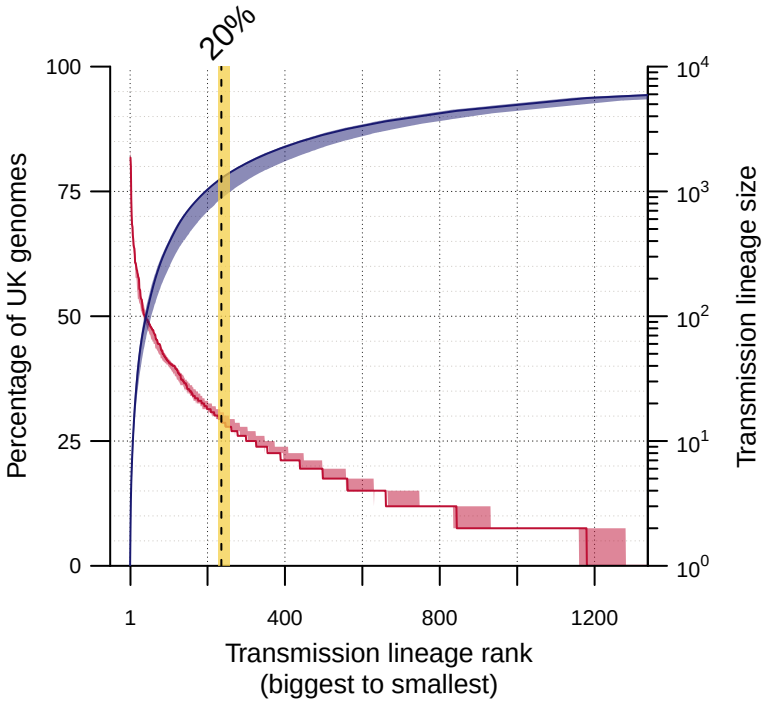

Figure 4: The percentage of UK genomes contained in transmission lineages, when ranked from biggest to smallest. The blue line and shading shows respectively the percentage of genomes in transmission lineages on the MCC trees and across the posterior tree distribution, respectively. The corresponding size of transmission lineages is shown in red on a logarithmic axis. The black dashed line and yellow shading show the 20% biggest transmission lineages on the MCC trees and across the posterior tree distribution, respectively.

posterior trees)

- The 20% biggest transmission lineages contain 77.59% of all UK genomes (in the MCC tree)
- The 8 biggest transmission lineages contain 26.04 [23.07, 28.51] of all UK genomes (across all 2000 posterior trees)
- The 8 biggest transmission lineages contain 26.35% of all UK genomes (in the MCC tree)

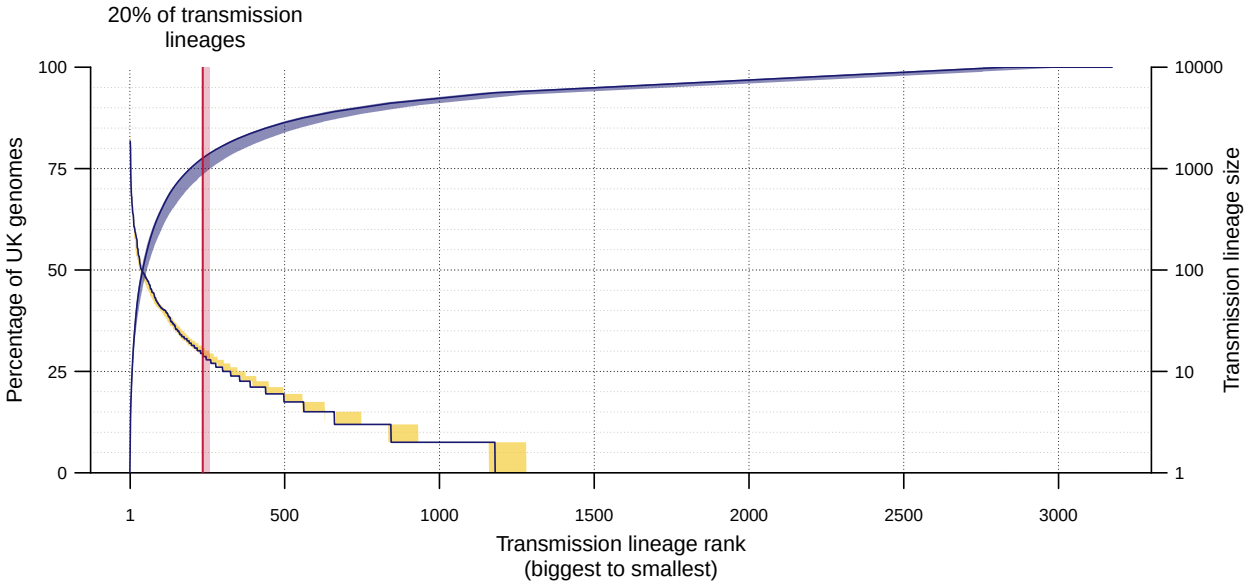

Figure 5: The percentage of UK genomes contained in transmission lineages, when ranked from biggest to smallest. The blue line and shading shows respectively the percentage of genomes in transmission lineages on the MCC trees and across the posterior tree distribution, respectively. The corresponding size of transmission lineages is shown in red on a logarithmic axis. The black dashed line and yellow shading show the 20% biggest transmission lineages on the MCC trees and across the posterior tree distribution, respectively.

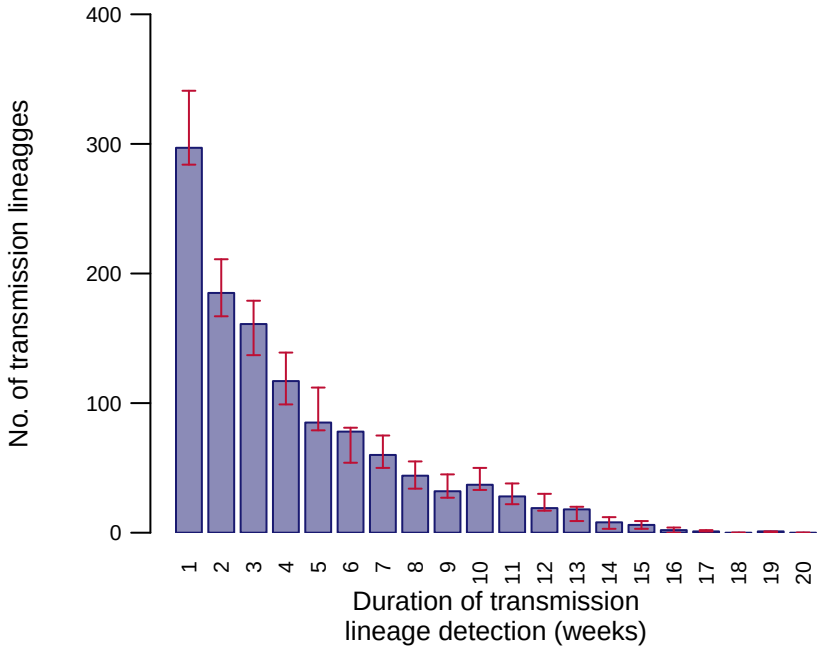

Figure 6: Distribution of UK transmission lineage sampling durations, aggregated by week. Blue bars show the number of transmission lineages that were observed over different durations in the MCC tree. Red bars show 95% HPD intervals for these numbers across the posterior tree distribution.

3 Transmission lineage TMRCA distribution

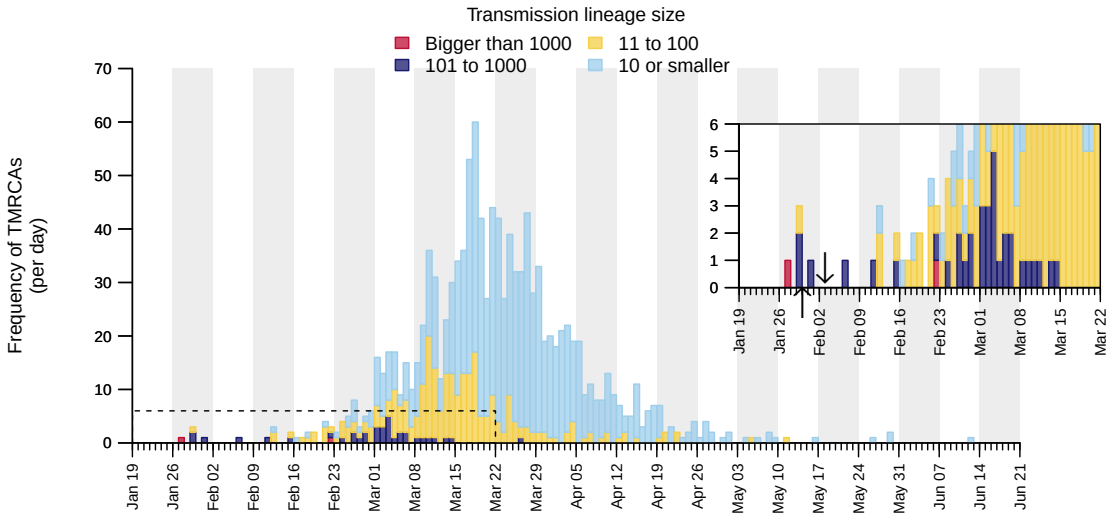

Figure 7: Histogram of lineage TMRCAs, coloured by lineage size. Inset: expanded view of the days prior to UK lockdown. Left-hand arrow = collection date of the UK's first laboratory-confirmed case; right-hand arrow = collection date of the earliest UK virus genome in our dataset.

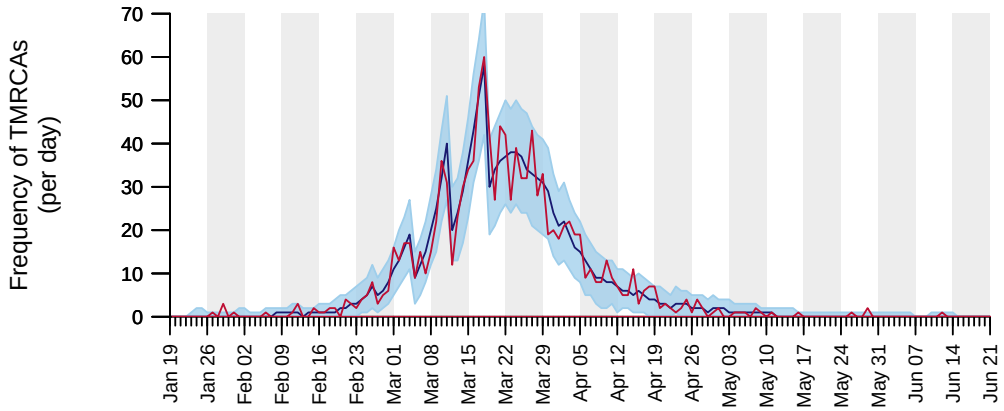

Figure 8: Comparison between the number of UK transmission lineage TMRCAs on each date in the MCC trees (red line) and across the 2000 posterior trees (median = blue line, 95% HPD interval = blue shading). Unevenness in this distribution is mostly likely caused by the phylogenetic constraints imposed by the sequence sampling times.

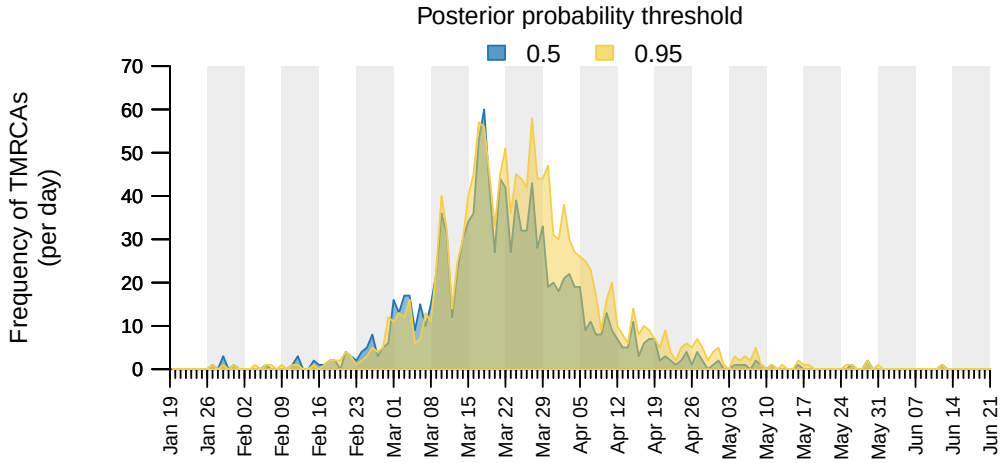

Figure 9: Comparison between TMRCA distributions with posterior probability thresholds of 0.5 and 0.95.

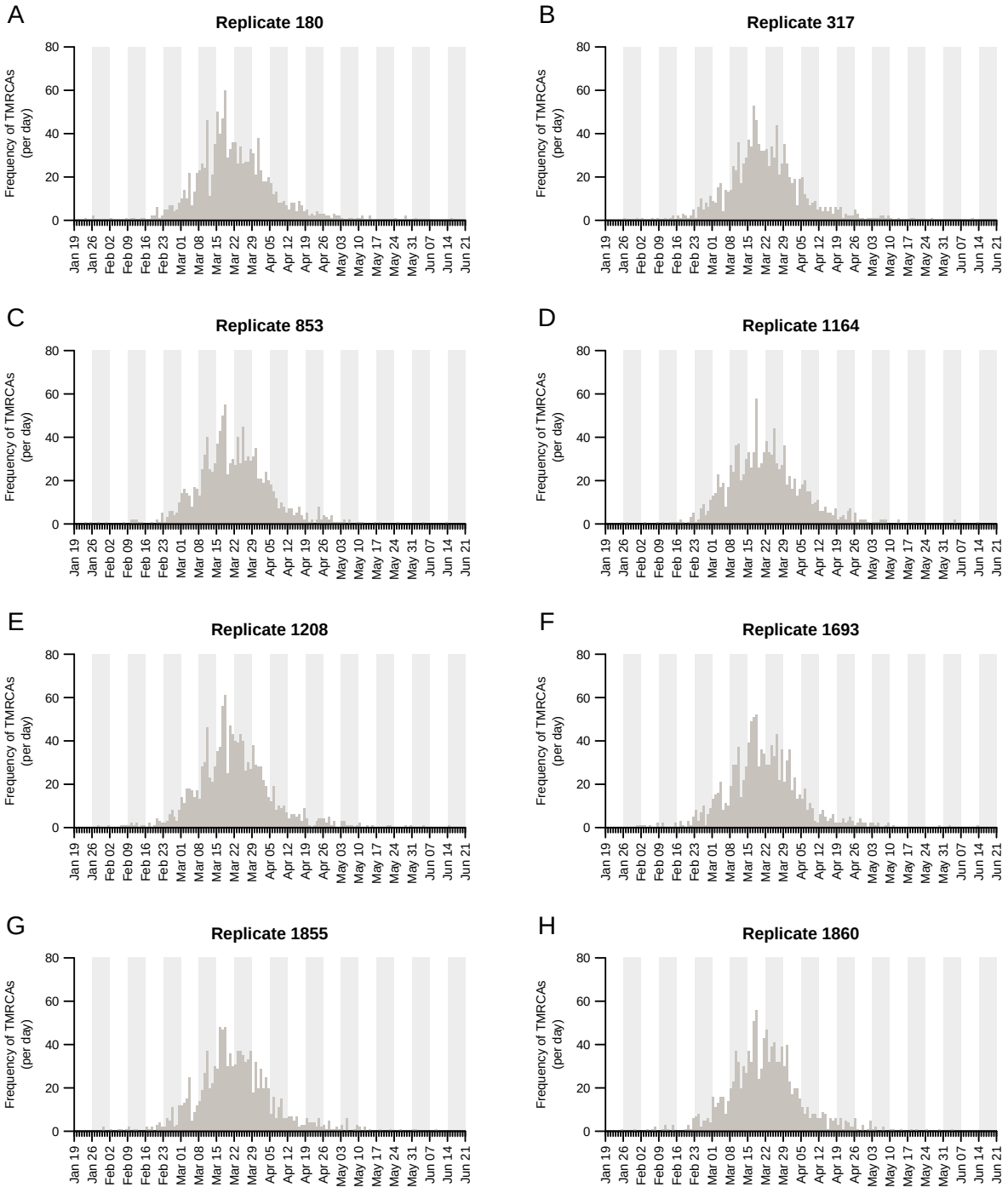

Figure 10: Histogram of lineage TRMCAs of 8 (of 2000) randomly selected posterior trees.

4 Size and duration vs. TMRCA

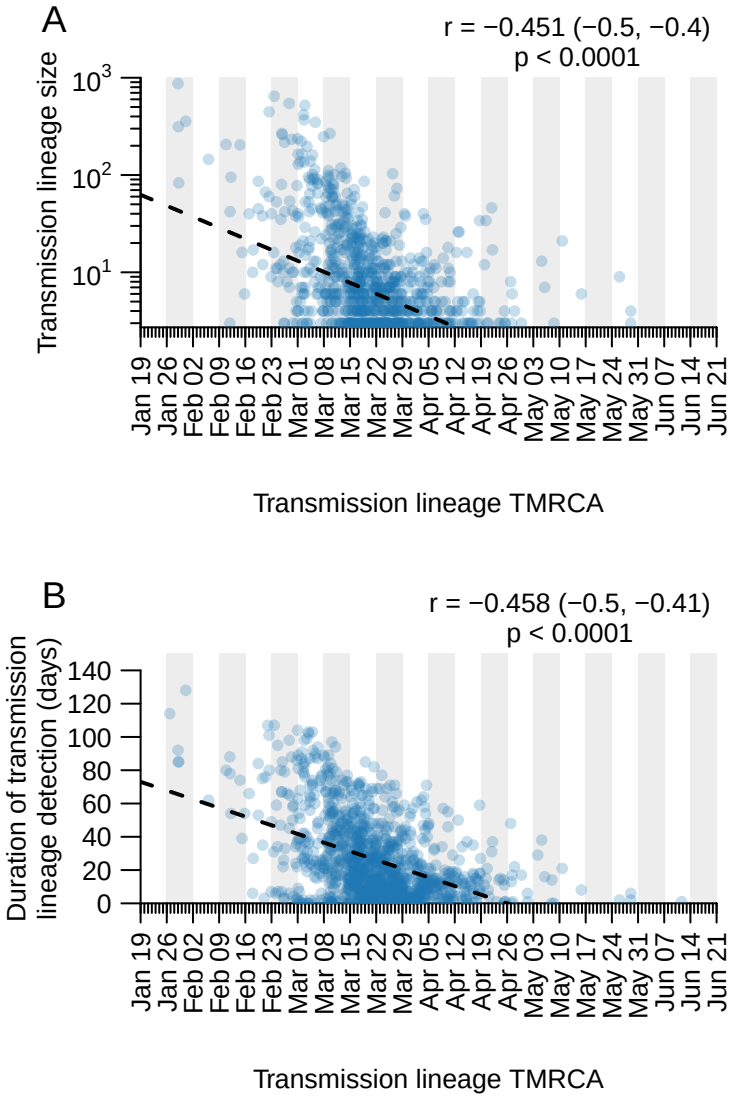

Figure 11: Scatterplots showing the relationship between (A) UK transmission lineage size and lineage TMRCA and between (B) UK transmission lineage sampling duration and lineage TMRCA. Pearson correlation coefficients, 95% CIs and p-values are shown in the top-right corners.

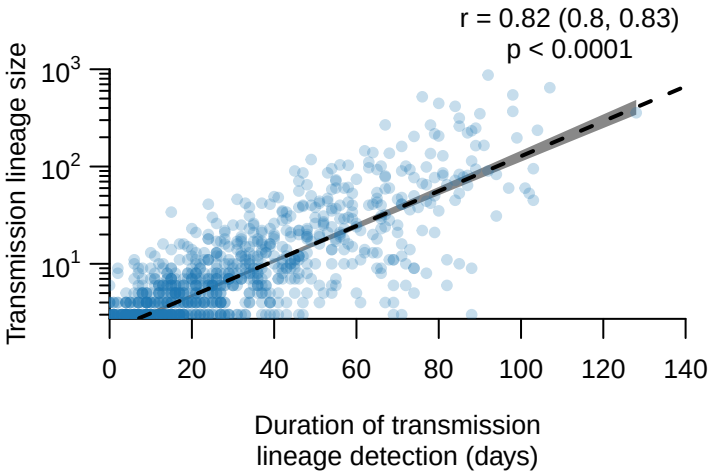

Figure 12: Scatterplot showing the strong relationship between UK transmission lineage size and sampling duration. The Pearson correlation coefficient, 95% CI and p-value are shown.

5 Time since lineage last sampled

```
## 19.292 sec elapsed
```

```
## 26961.12 sec elapsed
```

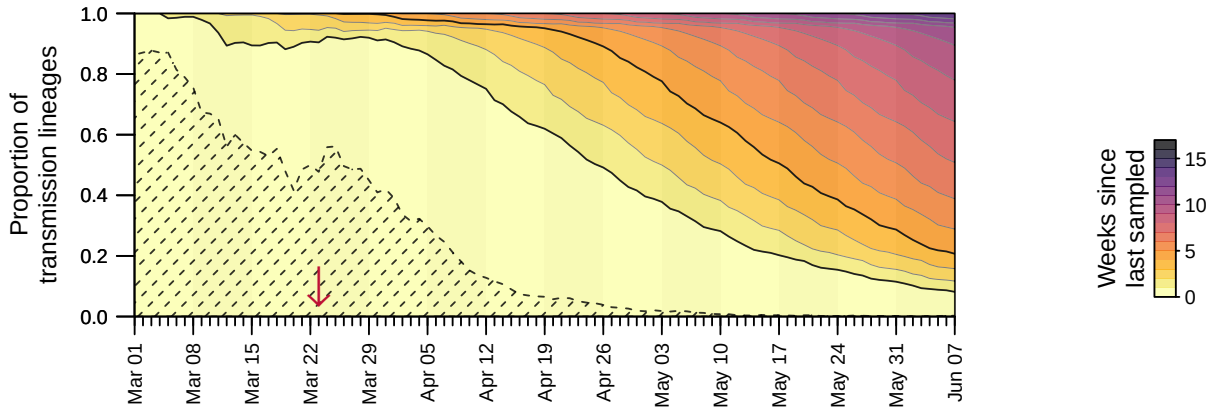

Figure 13: Trends through time in the detection of UK transmission lineages (proportions). For each day, all lineages detected up to that day are coloured by the time since the transmission lineage was last sampled. Isoclines correspond to weeks. Shaded area=transmission lineages that were first sampled <1 week ago. The red arrow indicates the start of the UK lockdown.

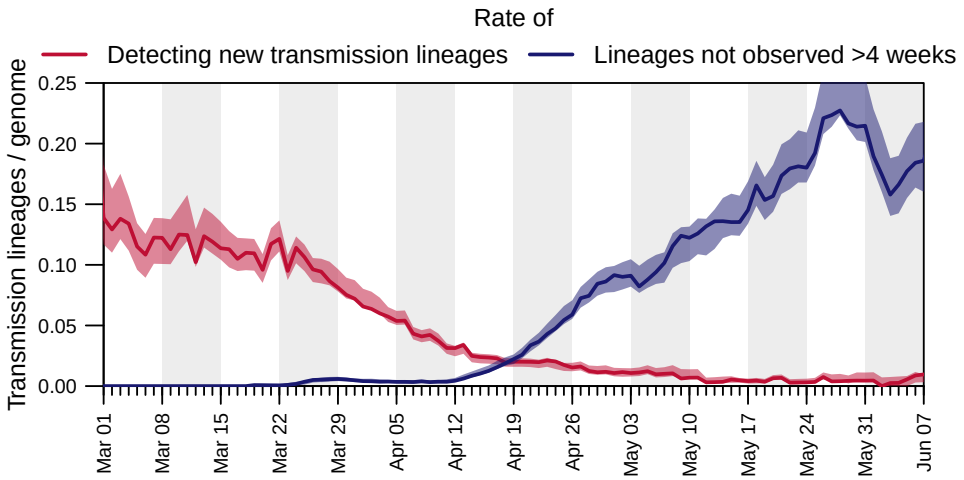

Figure 14: Red line=daily rate of detecting new transmission lineages. Blue line=rate at which lineages have not been observed for >4 weeks, shading=95% HPD across the posterior tree distribution.

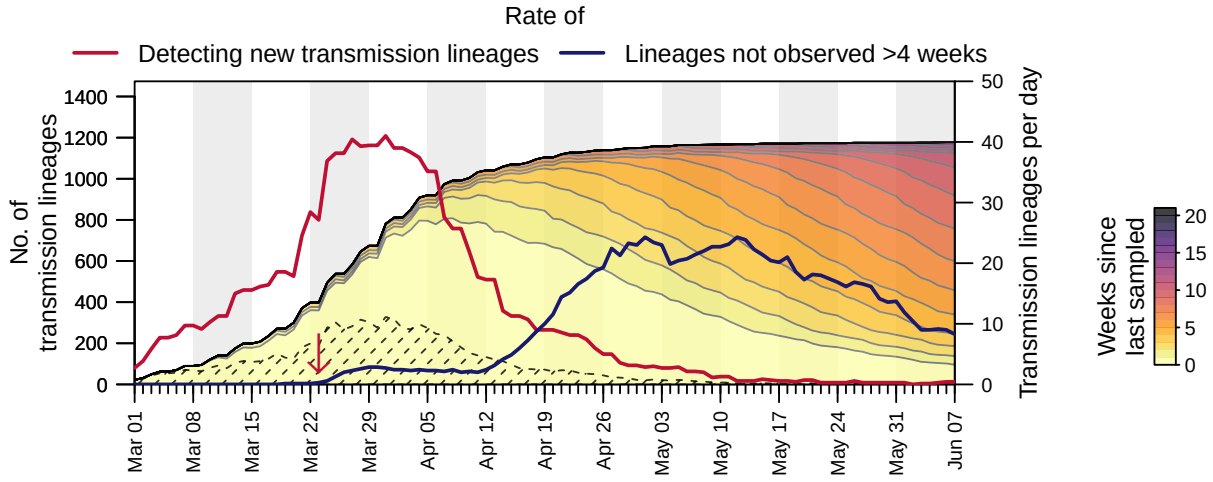

Figure 15: Trends through time in the detection of UK transmission lineages (absolute values). For each day, all lineages detected up to that day are coloured by the time since the transmission lineage was last sampled. Isoclines correspond to weeks. Shaded area=transmission lineages that were first sampled <1 week ago. The red arrow indicates the start of the UK lockdown.

6 Session info

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] tictoc_1.0        beastio_0.3.3     viridis_0.5.1     viridisLite_0.3.0
## [5] coda_0.19-3       treemap_2.4-2     gplots_3.0.1.1    lubridate_1.7.4
##
## loaded via a namespace (and not attached):
##  [1] boa_1.1.8-2        gtools_3.8.1       tidyselect_0.2.5   xfun_0.15
##  [5] purrr_0.3.3        lattice_0.20-38    colorspace_1.4-1   htmltools_0.4.0
##  [9] yaml_2.2.0         rlang_0.4.2        later_1.0.0        pillar_1.4.2
## [13] glue_1.3.1         RColorBrewer_1.1-2 lifecycle_0.1.0    stringr_1.4.0
## [17] munsell_0.5.0      gtable_0.3.0       caTools_1.17.1.3   evaluate_0.14
## [21] knitr_1.29         fastmap_1.0.1      httpuv_1.5.2       highr_0.8
## [25] Rcpp_1.0.3         KernSmooth_2.23-16 xtable_1.8-4       promises_1.1.0
## [29] scales_1.1.0       gdata_2.18.0       mime_0.7           gridExtra_2.3
## [33] ggplot2_3.2.1      digest_0.6.23      stringi_1.4.3      dplyr_0.8.3
## [37] shiny_1.4.0        grid_3.5.1         tools_3.5.1        bitops_1.0-6
## [41] magrittr_1.5       lazyeval_0.2.2     tibble_2.1.3       crayon_1.3.4
## [45] pkgconfig_2.0.3    gridBase_0.4-7     data.table_1.12.8  assertthat_0.2.1
## [49] rmarkdown_2.3      R6_2.4.1           igraph_1.2.4.2     compiler_3.5.1
```