COG-UK mini dataset exploratory analysis

Madeline Iseminger

20/01/2022

Introduction

This is an exploratory/remembering-how-to-use-R notebook, exploring a dataset of already-aligned ORW genomes from COG-UK.

```
metadata <- read.csv("/home/madeline/QIB_Internship/COG-UK-data/ORW_metadata.csv", header=TRUE)

head(metadata)
```

```
##   central_sample_id            sequence_name secondary_identifier sample_date
## 1       LIVE-E0239E England/LIVE-E0239E/2021                   NA  2021-02-14
## 2       NORW-22D223 England/NORW-22D223/2021                   NA  2021-02-10
## 3       NORW-22D1CC England/NORW-22D1CC/2021                   NA  2021-02-13
## 4       NORW-22D162 England/NORW-22D162/2021                   NA  2021-02-13
## 5       NORW-22B37D England/NORW-22B37D/2021                   NA  2021-01-09
## 6       NORW-22D2C9 England/NORW-22D2C9/2021                   NA  2021-02-10
##   epi_week country   adm1 is_surveillance is_travel_history travel_history
## 1       NA      UK UK-ENG               Y                NA             NA
## 2       NA      UK UK-ENG               Y                NA             NA
## 3       NA      UK UK-ENG               N                NA             NA
## 4       NA      UK UK-ENG               Y                NA             NA
## 5       NA      UK UK-ENG               Y                NA             NA
## 6       NA      UK UK-ENG               N                NA             NA
##   lineage lineage_support uk_lineage     del_lineage
## 1 B.1.1.7              NA     UK1476 del_trans_26865
## 2 B.1.1.7              NA     UK1476 del_trans_26865
## 3 B.1.1.7              NA     UK1364 del_trans_22142
## 4 B.1.1.7              NA     UK1476 del_trans_26865
## 5 B.1.1.7              NA      UK736 del_trans_26121
## 6 B.1.1.7              NA     UK1476 del_trans_26865
##                       phylotype
## 1 UK1476_1.33.13.70.46.1.48.1.1
## 2  UK1476_1.33.10.59.1.1.10.1.2
## 3                 UK1364_1.12.2
## 4       UK1476_1.33.10.59.1.1.1
## 5                     UK736_1.4
## 6   UK1476_1.33.13.265.18.1.1.4
```

Exploration

Total counts over all time for each lineage:

Var1	Freq
	47
B	76
B.1	137
B.1.1.1	20
B.1.1.10	8
B.1.1.101	1
B.1.1.102	2
B.1.1.114	9
B.1.1.115	1
B.1.1.119	254
B.1.1.127	28
B.1.1.130	9
B.1.1.147	1
B.1.1.159	1
B.1.1.164	1
B.1.1.170	20
B.1.1.175	2
B.1.1.182	4
B.1.1.183	10
B.1.1.184	2
B.1.1.189	5
B.1.1.198	199
B.1.1.208	2
B.1.1.211	1
B.1.1.215	9
B.1.1.220	1
B.1.1.240	3
B.1.1.251	3
B.1.1.269	16
B.1.1.274	2
B.1.1.279	112
B.1.1.286	1
B.1.1.3	14
B.1.1.305	1
B.1.1.306	1
B.1.1.307	3
B.1.1.311	4
B.1.1.314	1
B.1.1.315	11
B.1.1.37	15
B.1.1.38	3
B.1.1.4	7
B.1.1.41	2
B.1.1.51	1
B.1.1.64	21
B.1.1.7	1263
B.1.1.74	1
B.1.1.88	1
B.1.105	6
B.1.111	10
B.1.13	1
B.1.146	2

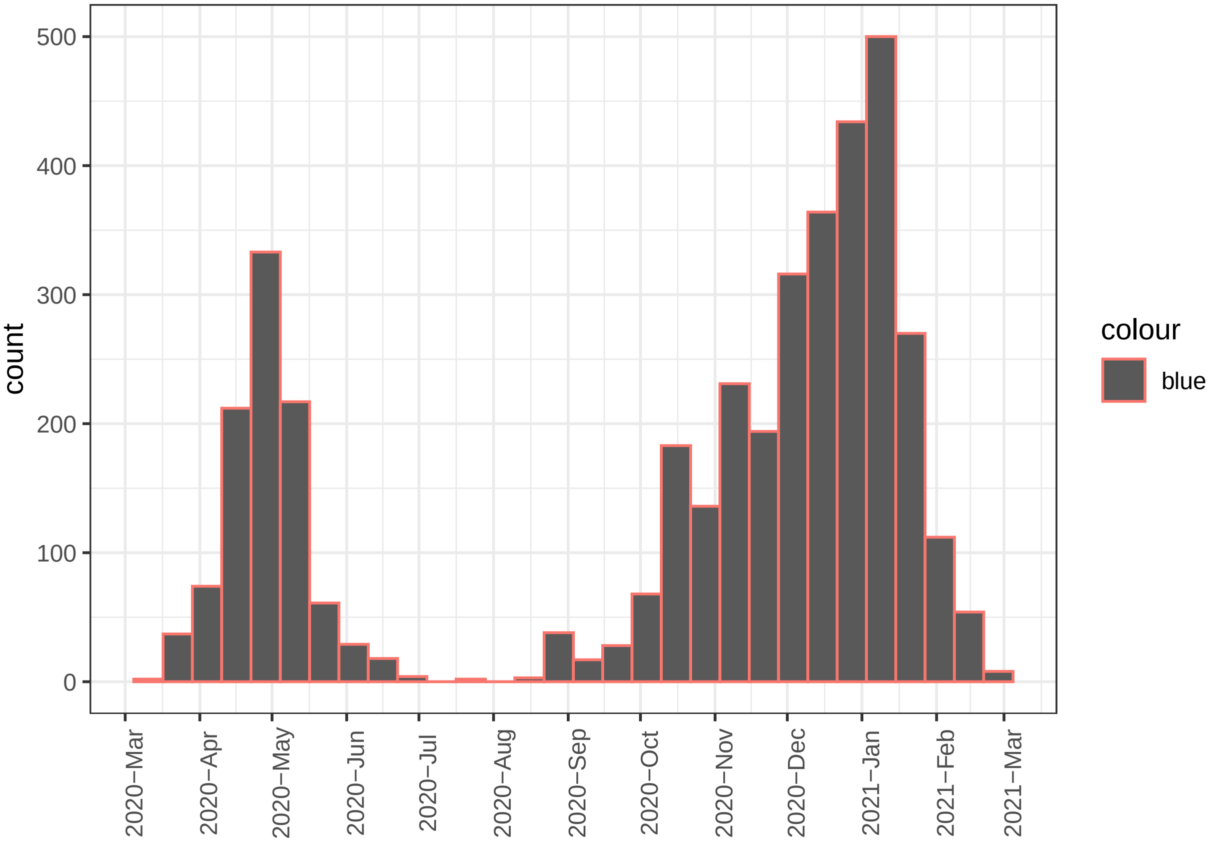
Var1	Freq
B.1.160	10
B.1.160.7	1
B.1.177	926
B.1.177.10	4
B.1.177.11	1
B.1.177.13	5
B.1.177.16	31
B.1.177.17	2
B.1.177.18	7
B.1.177.19	5
B.1.177.20	5
B.1.177.26	100
B.1.177.4	20
B.1.177.5	1
B.1.177.6	10
B.1.177.7	1
B.1.177.8	5
B.1.177.9	1
B.1.201	85
B.1.218	2
B.1.221	3
B.1.221.1	6
B.1.222	1
B.1.225	66
B.1.235	3
B.1.236	3
B.1.250	2
B.1.258	19
B.1.258.4	1
B.1.258.6	4
B.1.351	3
B.1.36	3
B.1.36.1	75
B.1.36.17	27
B.1.36.28	1
B.1.36.9	5
B.1.389	1
B.1.391	5
B.1.392	86
B.1.408	2
B.1.523	4
B.1.88	1
B.1.93	23
B.1.98	3
B.27	1
B.28	2
B.29	2
B.3	10
B.39	2
B.40	9
B.48	6
B.52	1

Var1	Freq
H.1	34
P.2	1

Check how many sequences there are over time:

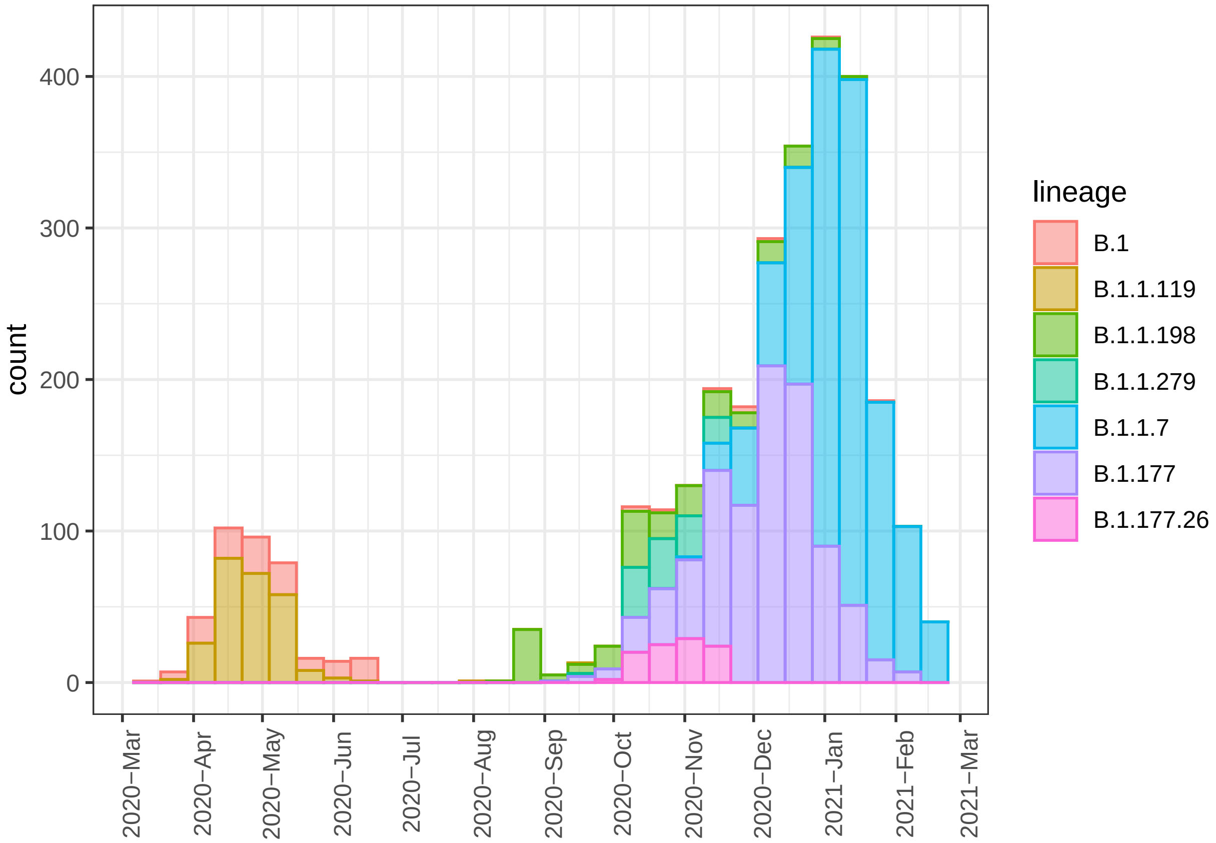
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 47 rows containing non-finite values (stat_bin).
```

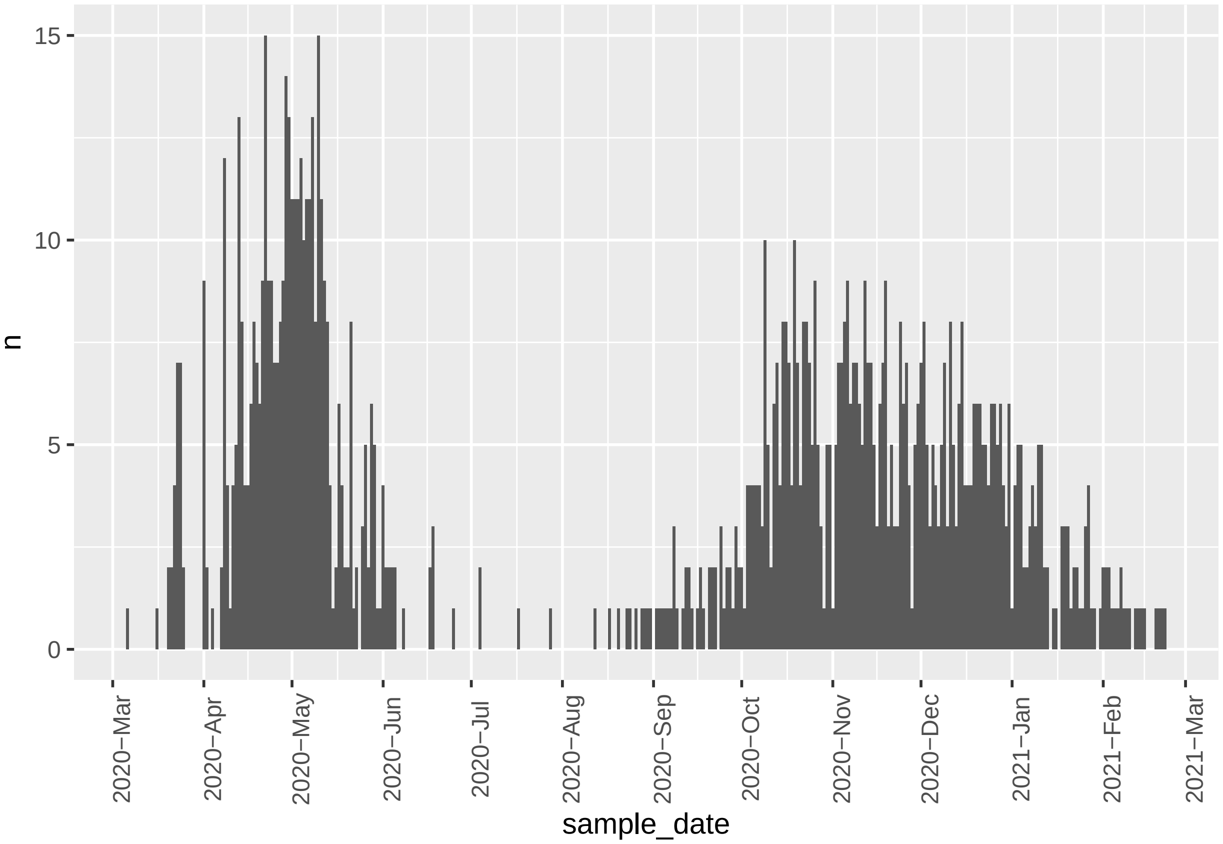


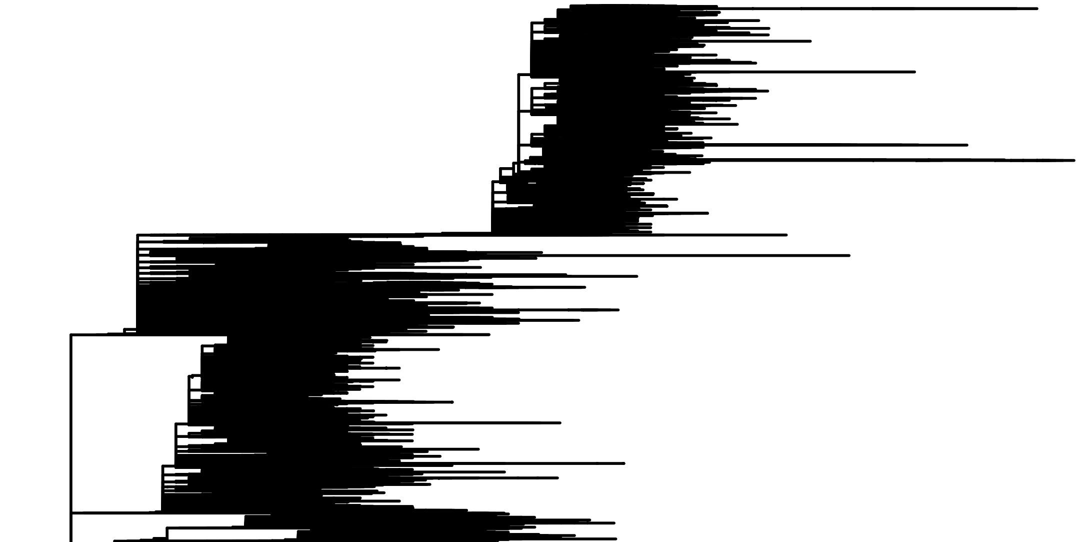
Now look at all the lineages with >=100 sequences present in the dataset. Plot a histogram by date with only these sequences.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Count the number of unique lineages per day:

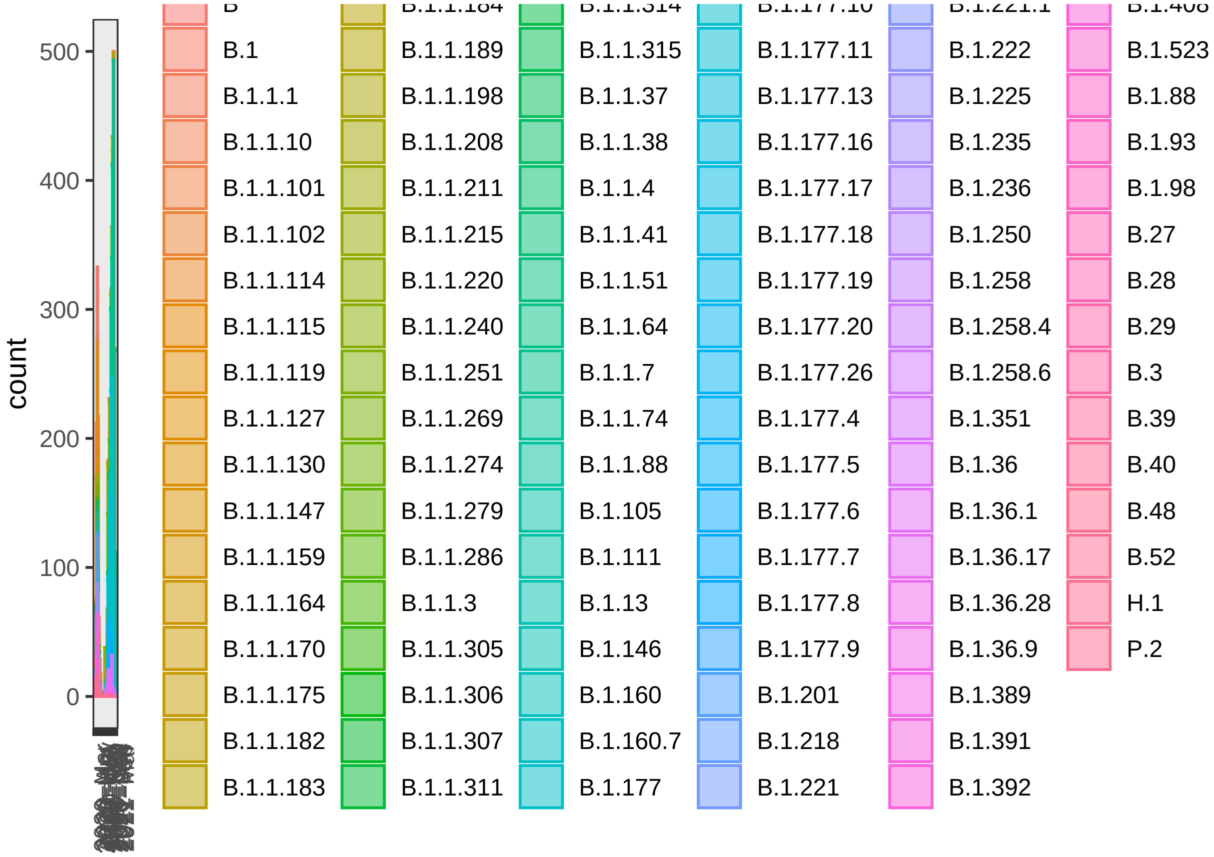


Read phylogenetic tree: 

Plot the total counts of each lineage per time unit:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 47 rows containing non-finite values (stat_bin).
```


count
500
400
300
200
100
0
B
B.1
B.1.1.1
B.1.1.10
B.1.1.101
B.1.1.102
B.1.1.114
B.1.1.115
B.1.1.119
B.1.1.127
B.1.1.130
B.1.1.147
B.1.1.159
B.1.1.164
B.1.1.170
B.1.1.175
B.1.1.182
B.1.1.183
B.1.1.184
B.1.1.189
B.1.1.198
B.1.1.208
B.1.1.211
B.1.1.215
B.1.1.220
B.1.1.240
B.1.1.251
B.1.1.269
B.1.1.274
B.1.1.279
B.1.1.286
B.1.1.3
B.1.1.305
B.1.1.306
B.1.1.307
B.1.1.311
B.1.1.314
B.1.1.315
B.1.1.37
B.1.1.38
B.1.1.4
B.1.1.41
B.1.1.51
B.1.1.64
B.1.1.7
B.1.1.74
B.1.1.88
B.1.105
B.1.111
B.1.13
B.1.146
B.1.160
B.1.160.7
B.1.177
B.1.177.10
B.1.177.11
B.1.177.13
B.1.177.16
B.1.177.17
B.1.177.18
B.1.177.19
B.1.177.20
B.1.177.26
B.1.177.4
B.1.177.5
B.1.177.6
B.1.177.7
B.1.177.8
B.1.177.9
B.1.201
B.1.218
B.1.221
B.1.221.1
B.1.222
B.1.225
B.1.235
B.1.236
B.1.250
B.1.258
B.1.258.4
B.1.258.6
B.1.351
B.1.36
B.1.36.1
B.1.36.17
B.1.36.28
B.1.36.9
B.1.389
B.1.391
B.1.392
B.1.408
B.1.523
B.1.88
B.1.93
B.1.98
B.27
B.28
B.29
B.3
B.39
B.40
B.48
B.52
H.1
P.2