EBOV sequence datasets

Louis du Plessis

October 31, 2018

Contents

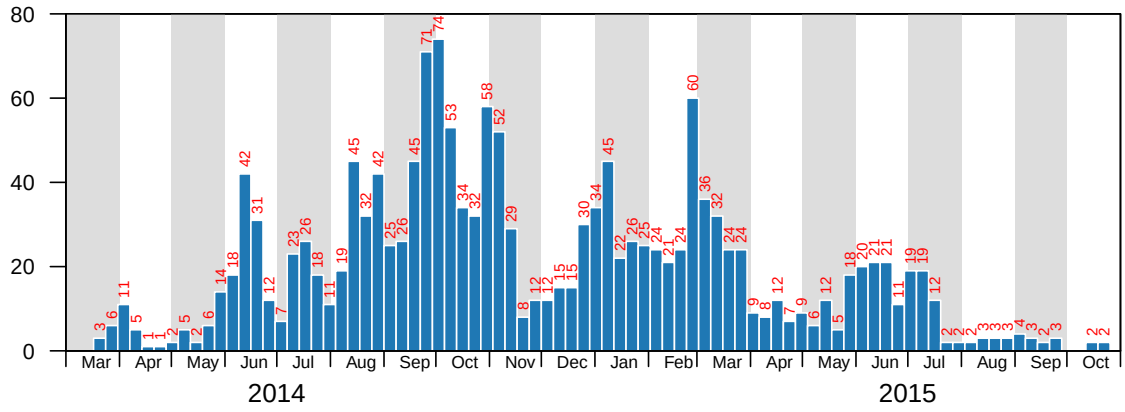
1 Summary

Get a few smaller subsets of the dataset on https://github.com/ebov/space-time.

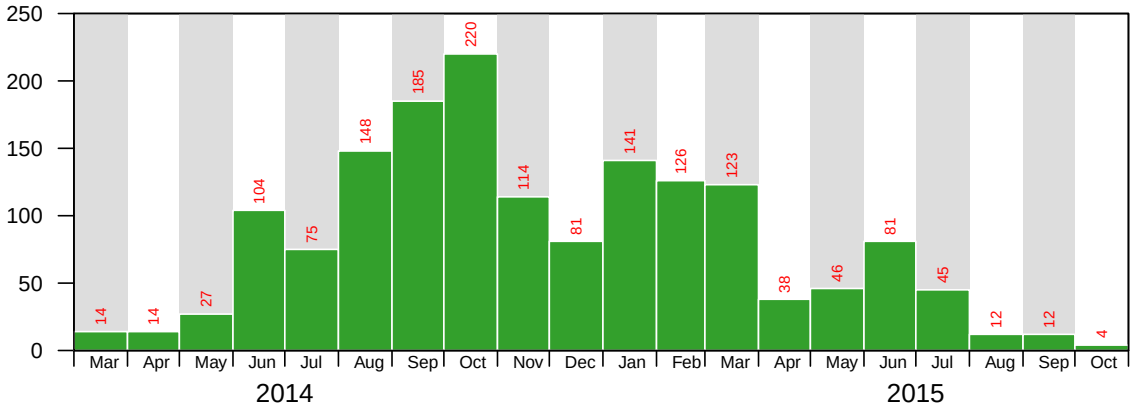
---

2 Datasets

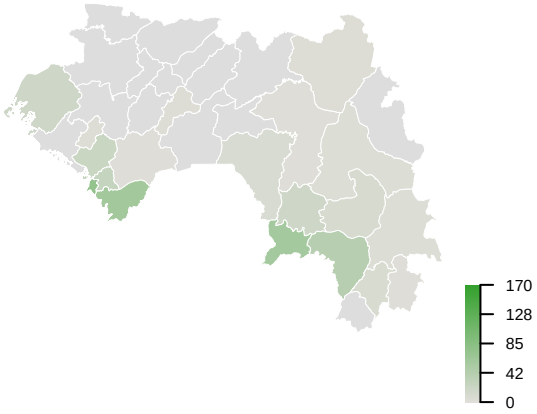
2.1 Complete dataset

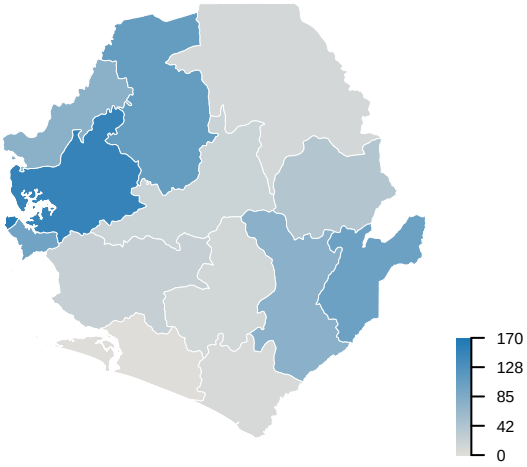


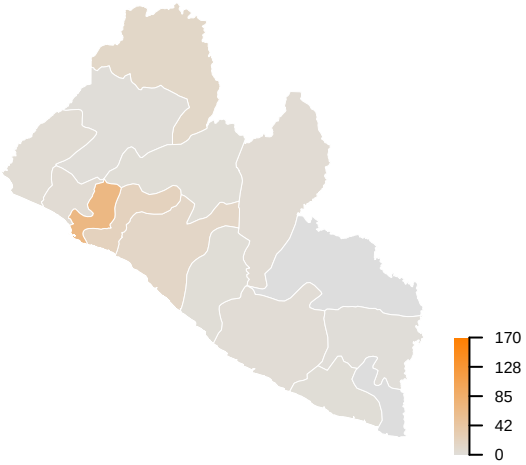
Complete dataset (weekly counts)
(1610 sequences)

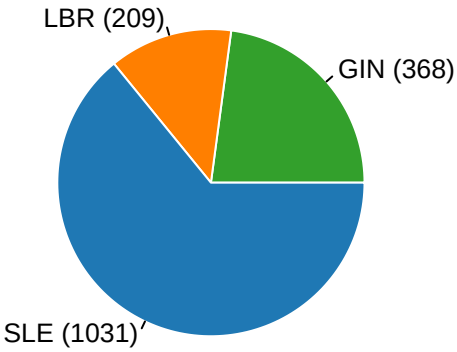


Complete dataset (monthly counts)
(1610 sequences)

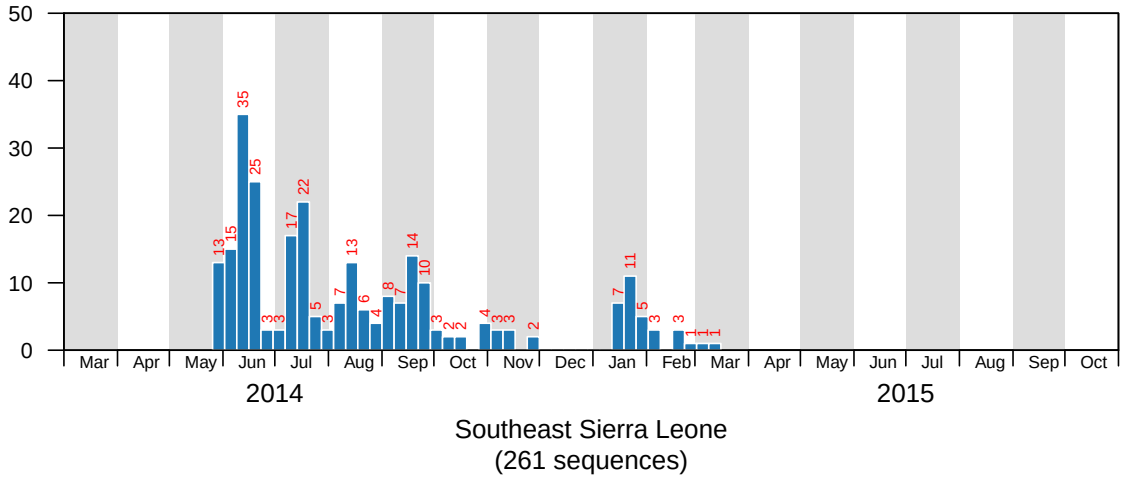

170
128
85
42
0


170
128
85
42
0

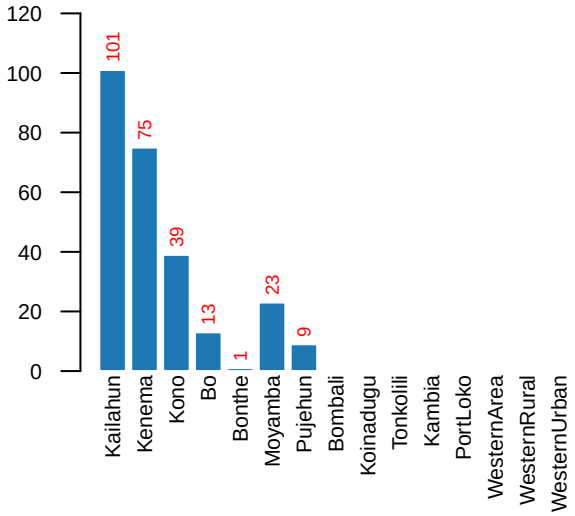

170
128
85
42
0

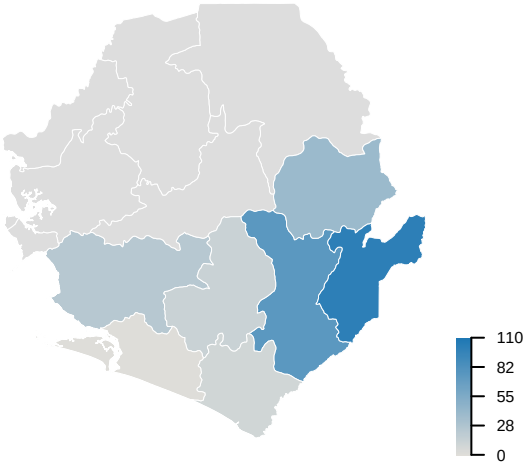

LBR (209)
GIN (368)
SLE (1031)

2.2 Southeast Sierra Leone

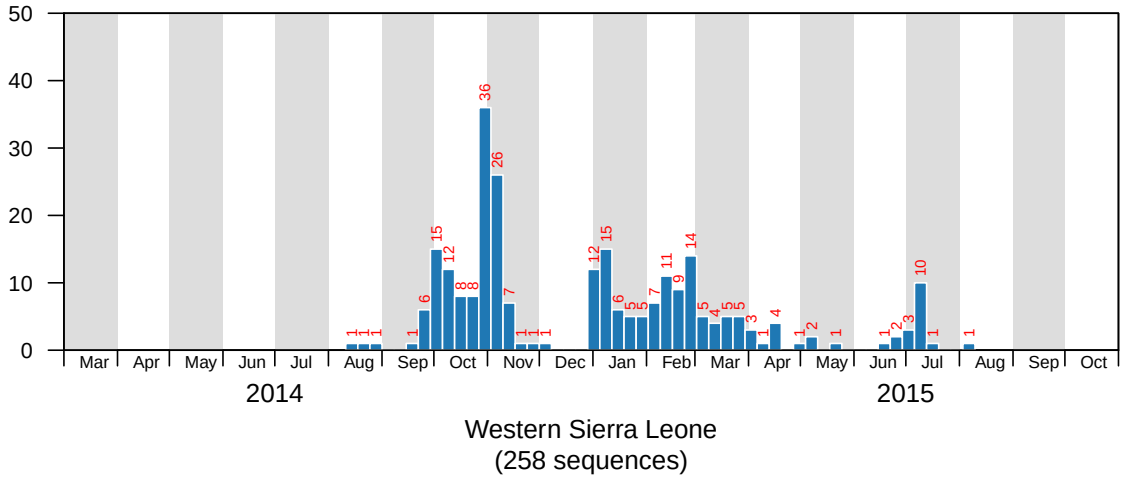


Southeast Sierra Leone
(261 sequences)

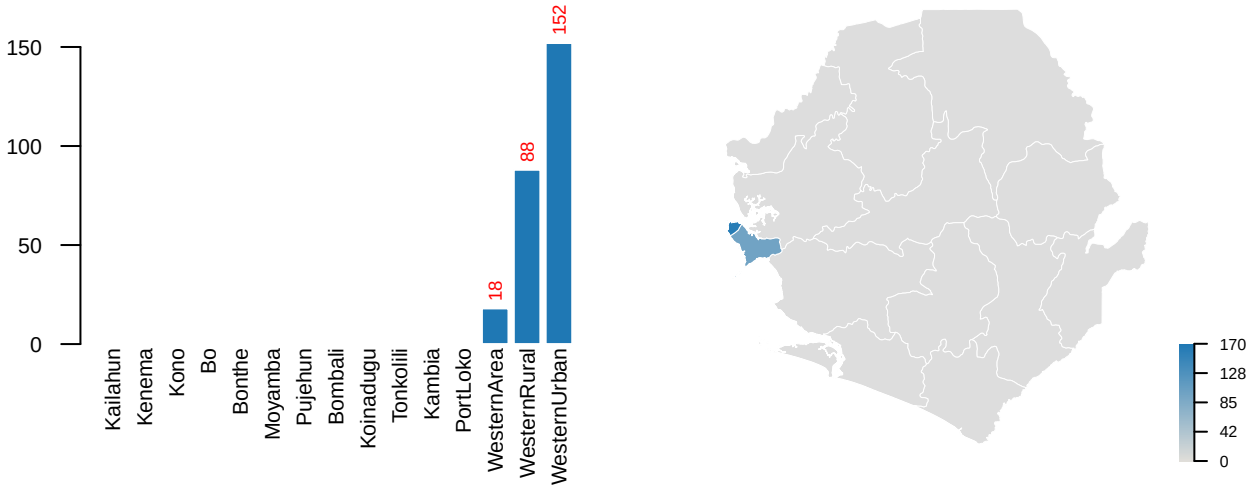




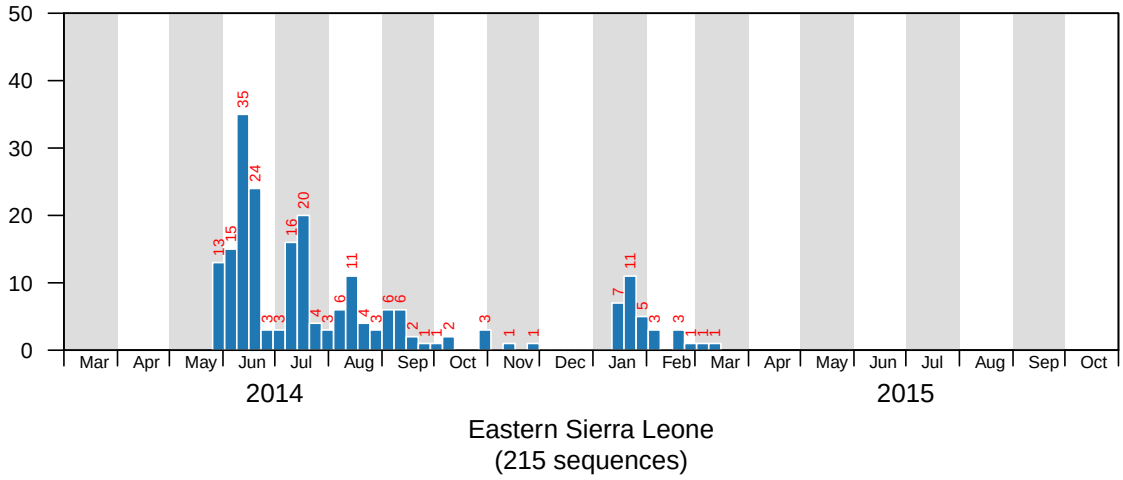
2.3 Western Sierra Leone



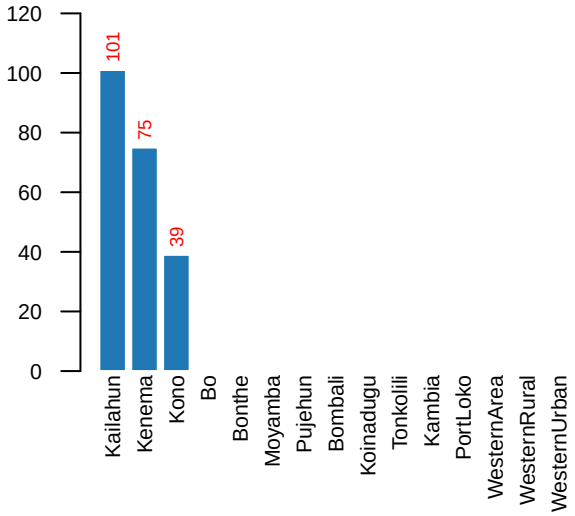
Western Sierra Leone
(258 sequences)

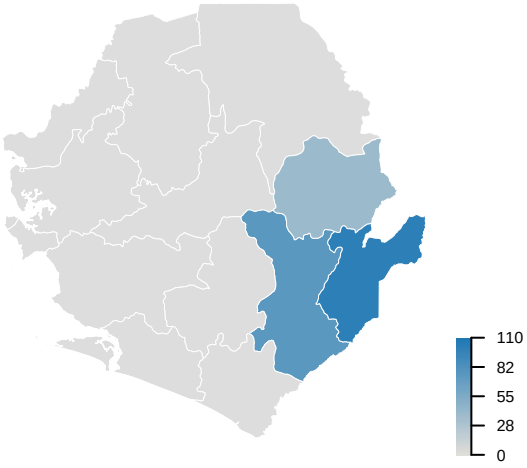


2.4 Eastern Sierra Leone

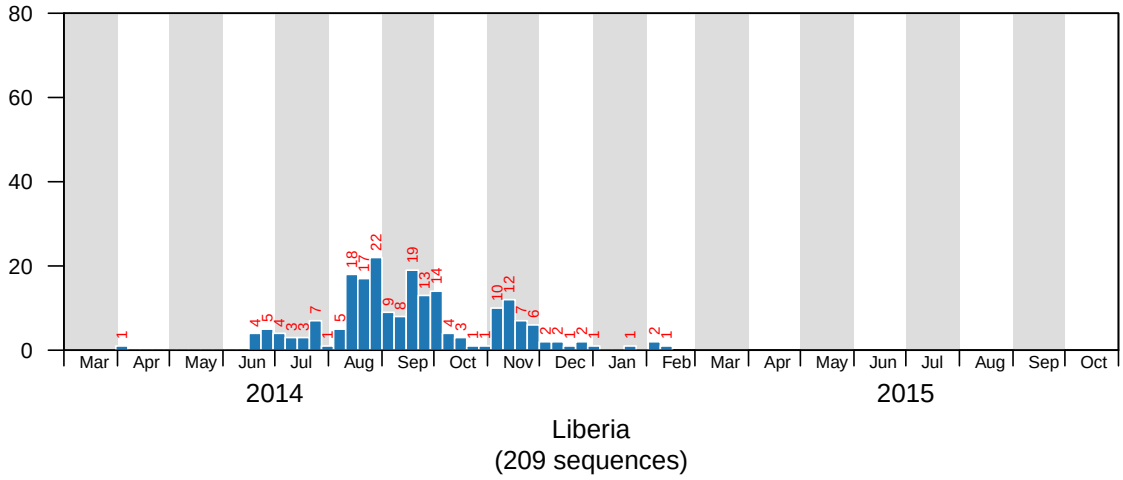


Eastern Sierra Leone
(215 sequences)

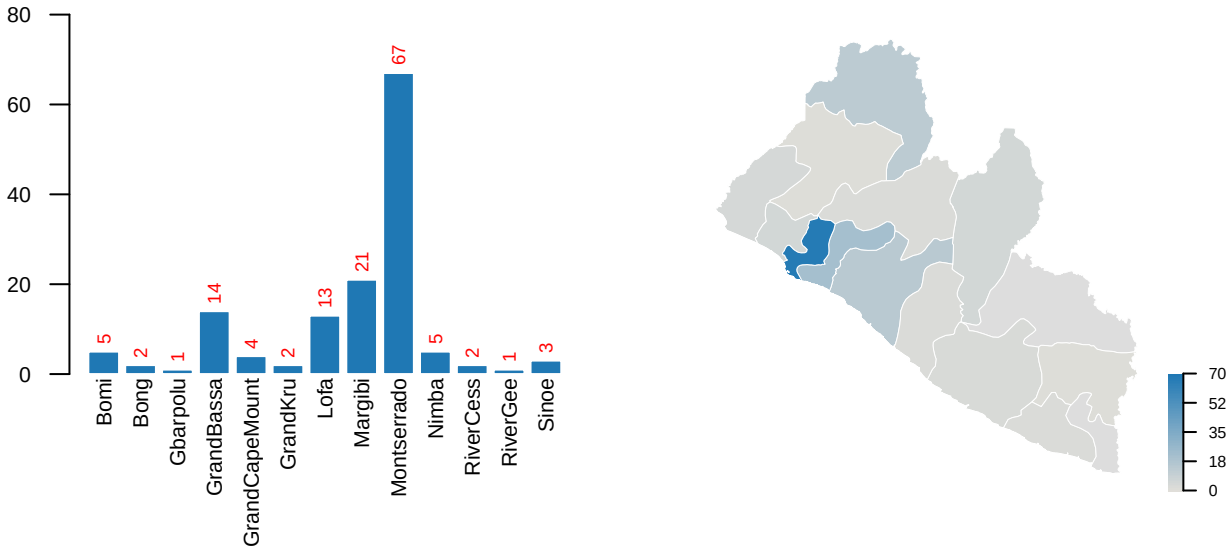




2.5 Liberia

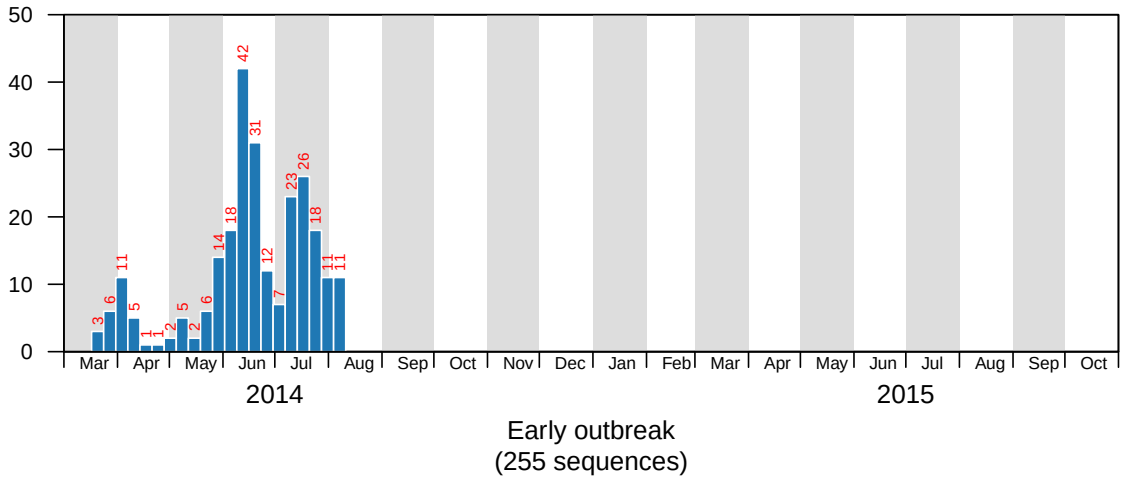


Liberia
(209 sequences)

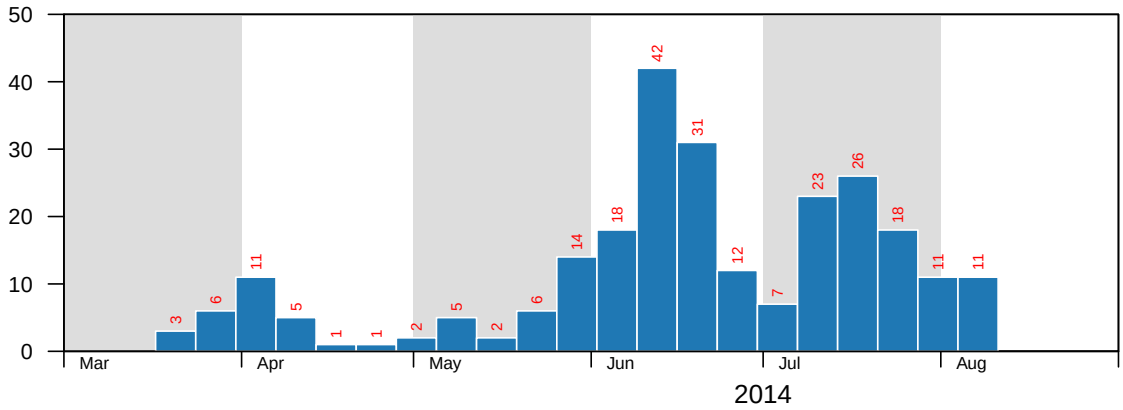


2.6 Early outbreak

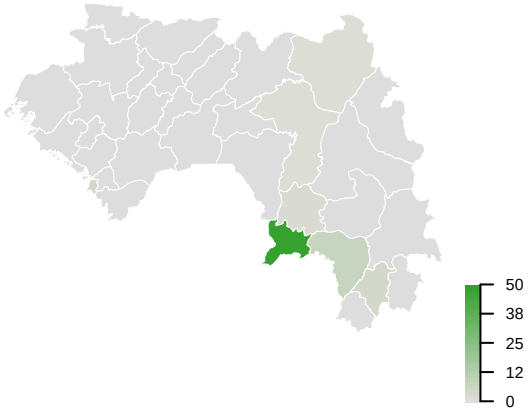
All sequences until 8 August (WHO declares public health emergency).

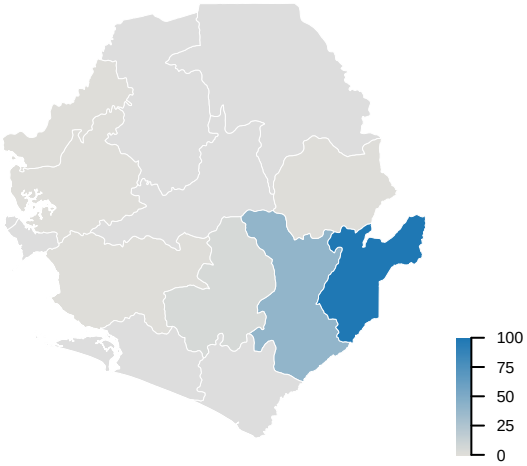


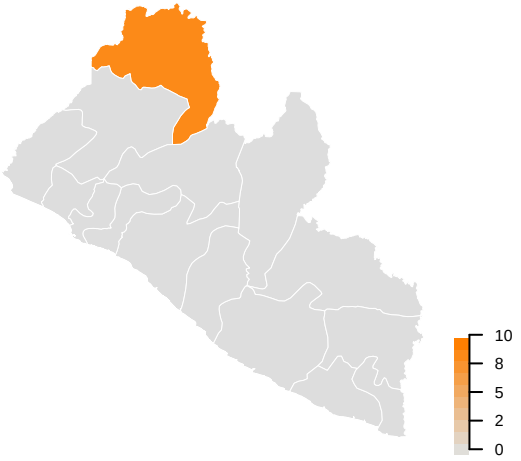
Early outbreak
(255 sequences)

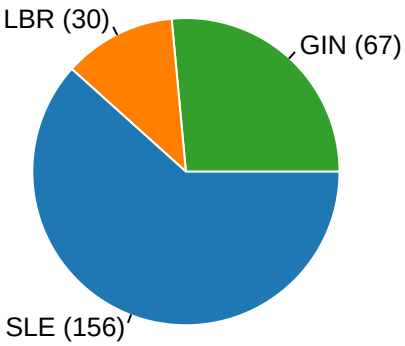


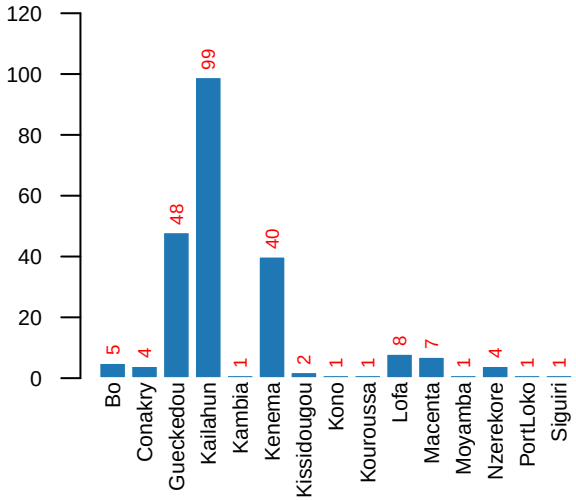
Early outbreak
(255 sequences)


50
38
25
12
0


100
75
50
25
0

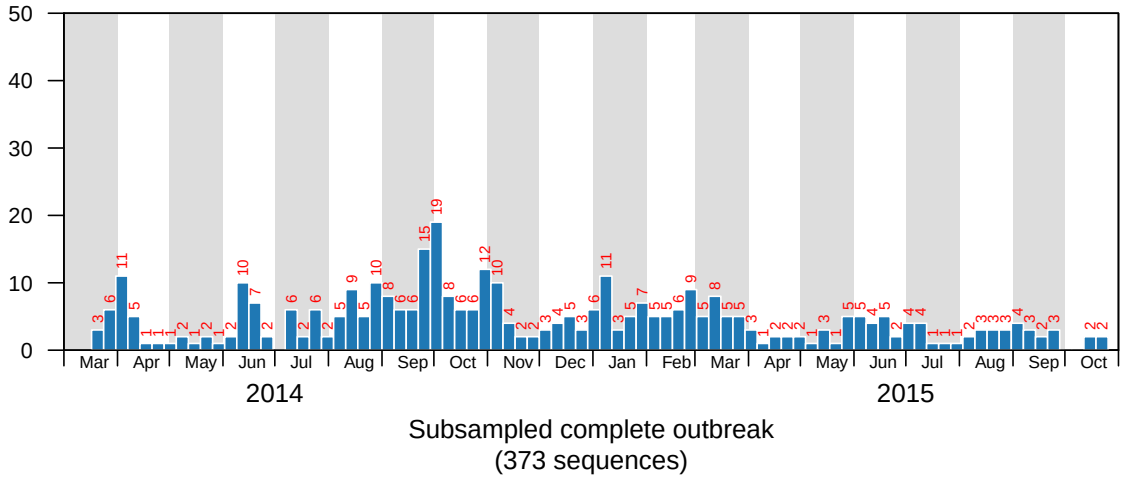

10
8
5
2
0


LBR (30)
GIN (67)
SLE (156)

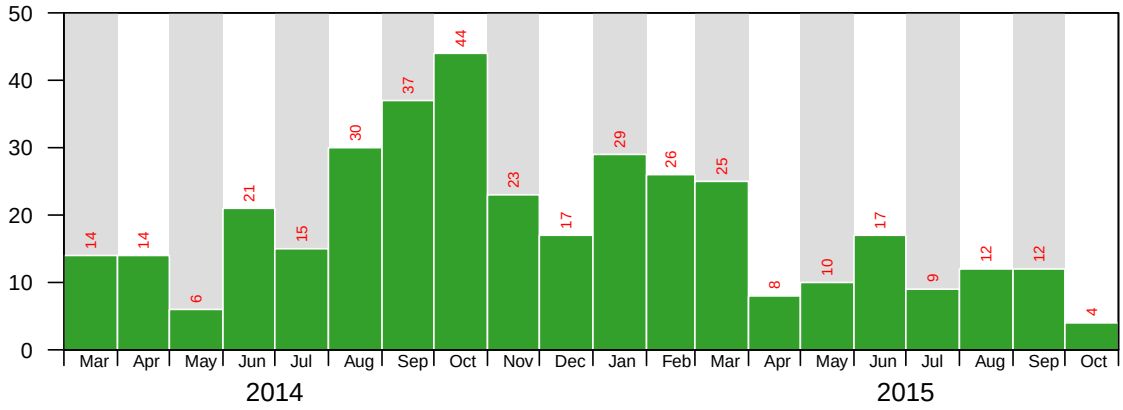

120
100
80
60
40
20
0
5
4
48
99
1
40
2
1
1
8
7
1
4
1
1
Bo
Conakry
Gueckedou
Kailahun
Kambia
Kenema
Kissidougou
Kono
Kouroussa
Lofa
Macenta
Moyamba
Nzerekore
PortLoko
Siguiri

2.7 Subsampled complete outbreak

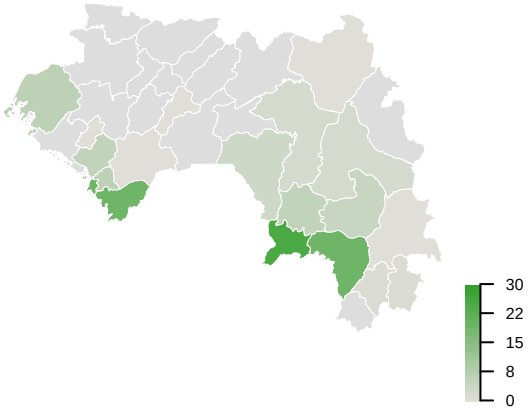
Complete outbreak subsampled to approximately a quarter of sequences and roughly stratified by month.

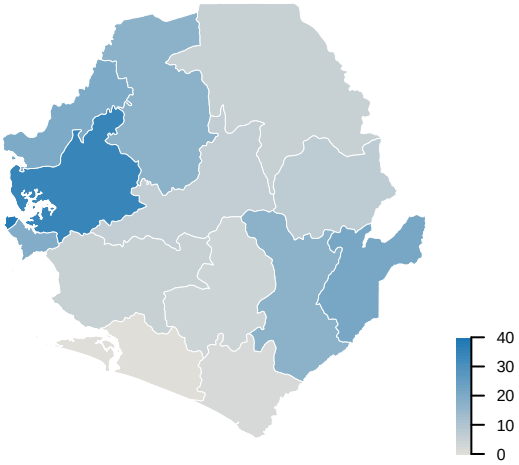


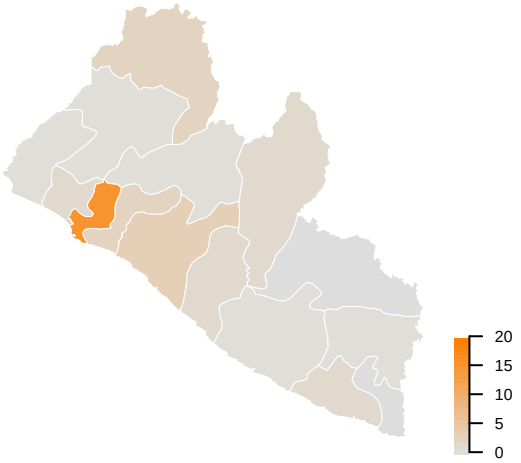
Subsampled complete outbreak
(373 sequences)

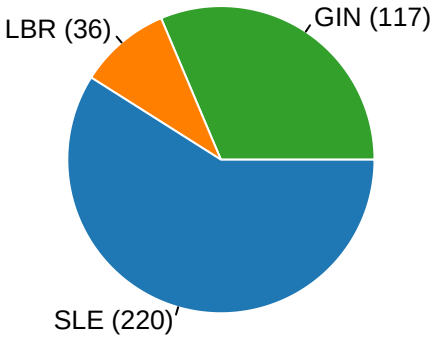


Subsampled complete outbreak
(373 sequences)


30
22
15
8
0

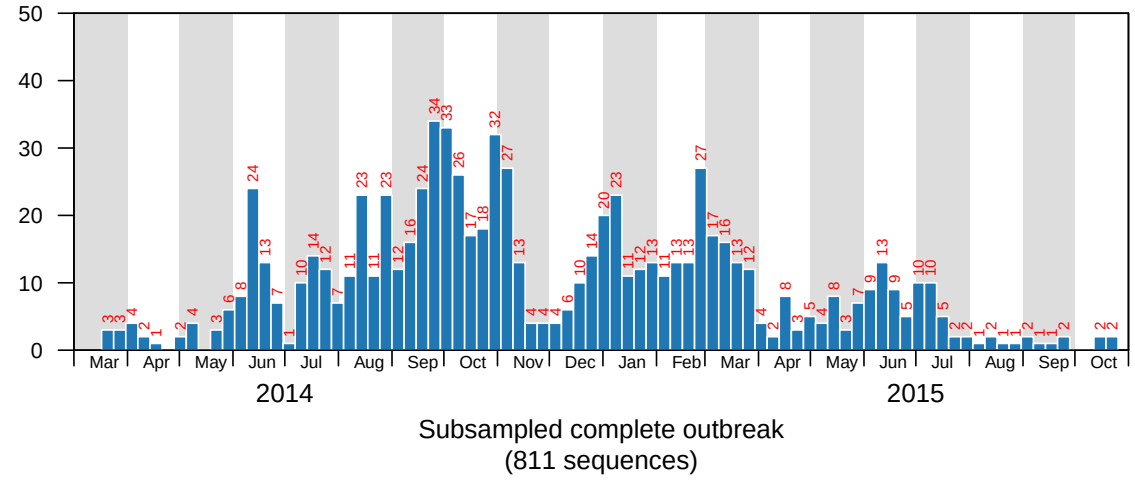

40
30
20
10
0


20
15
10
5
0

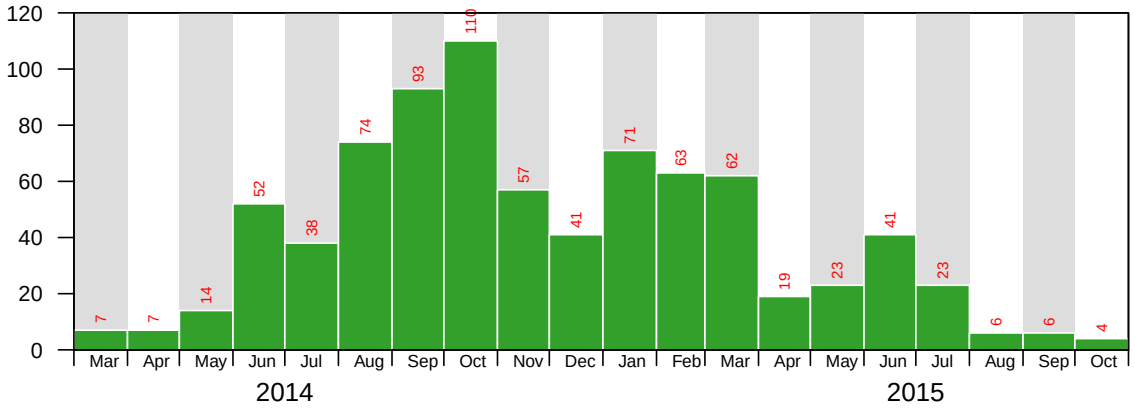

LBR (36)
GIN (117)
SLE (220)

2.8 Subsampled complete outbreak big

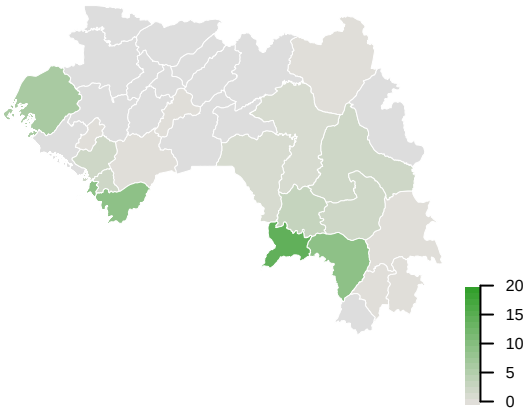
Complete outbreak subsampled to approximately half of all sequences and roughly stratified by month.

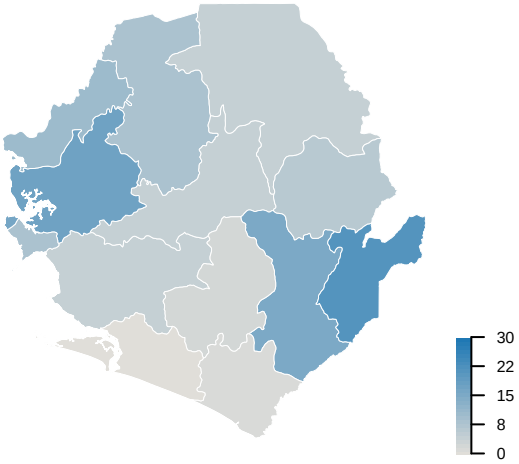


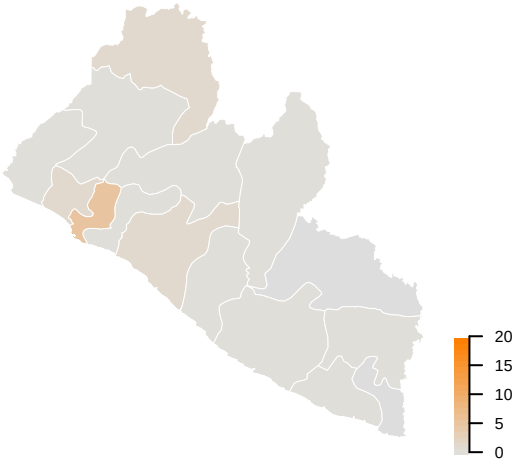
Subsampled complete outbreak
(811 sequences)

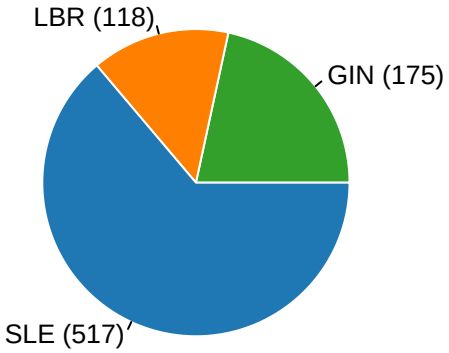


Subsampled complete outbreak
(811 sequences)









3 Dataset statistics

Table 1: Dataset statistics

names	sequences	earliest	latest	range
complete	1610	2014-03-17	2015-10-24	586
region-sle-southeast	261	2014-05-25	2015-03-10	289
region-sle-west	258	2014-08-15	2015-08-06	356
region-sle-east	215	2014-05-25	2015-03-10	289
country-lbr	209	2014-04-01	2015-02-14	319
special-early	255	2014-03-17	2014-08-05	141
special-suball	373	2014-03-17	2015-10-24	586
special-subbig	811	2014-03-17	2015-10-24	586

4 Select sequences from Fasta file

```
# Load Conda environment
source /Users/user/anaconda3/envs/bio/bin/activate bio

# Extract Others
for i in `ls ../results/datasets/*.csv`
do
    # Full-length genomes
    python ../scripts/selectsequences.py -i ${i} -a ../data/sequence/Data/Makona_1610_genomes_2016-0

    # Coding sequences
    python ../scripts/selectsequences.py -i ${i} -a ../results/datasets/Makona_1610_cds.trimmed.fas

    # Noncoding sequences
    python ../scripts/selectsequences.py -i ${i} -a ../results/datasets/Makona_1610_ig.trimmed.fas -
done
```