# Lineage Growth Rate EDA

SARS-CoV-2 Lineages

The first date of the MA genomic data is 2020-01-29 and the last date is 2023-01-31. The number of unique SAR-CoV-2 lineages during this period is 693.

Reporting lags

Each sequenced sample has a date the sample was taken and the date the sample was submitted. The reporting lag for each sample is the time elapsed between those two dates. Our training period starts at the beginning of March 2021, in which the median lags seem reasonable.
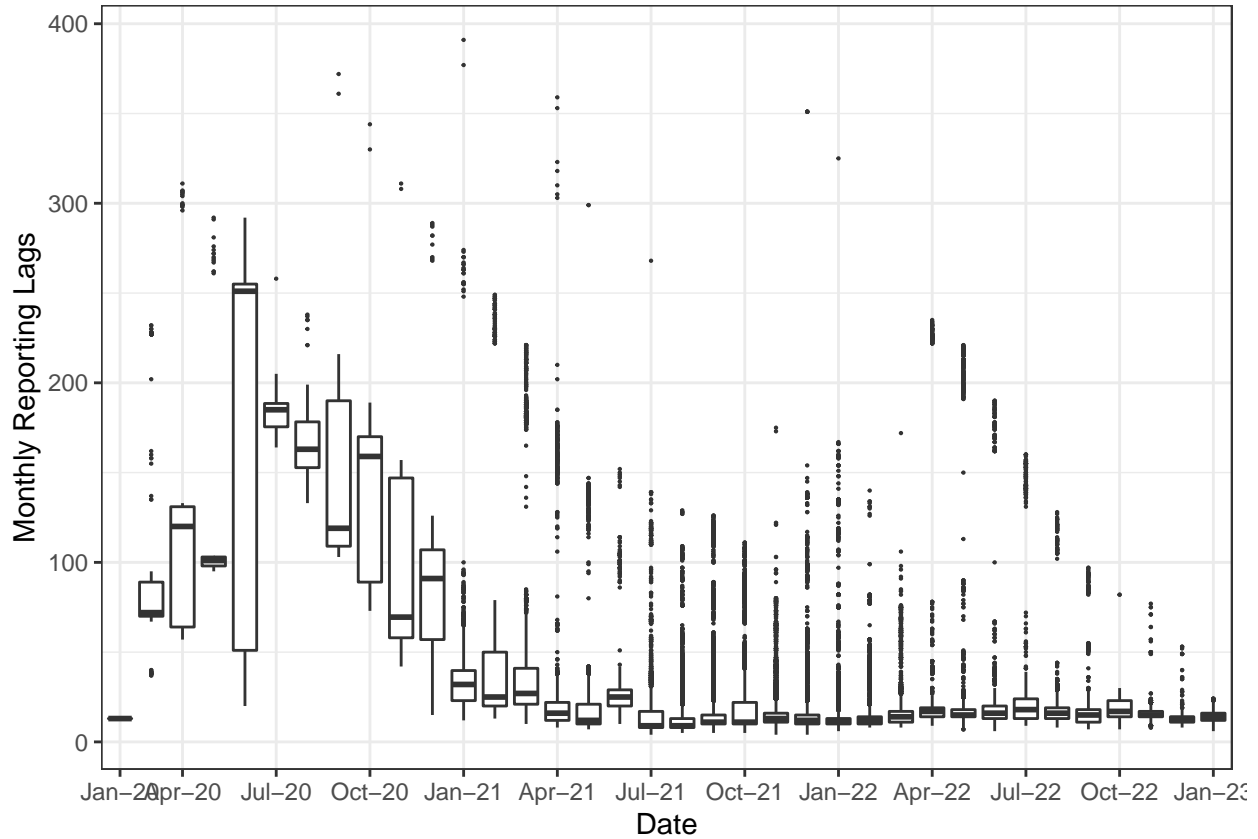
Number of lineages

There are 693 unique lineages in the data, below are the monthly boxplot number of unique lineages on a given date. There were a large number of lineages circulating in the later half of 2021.
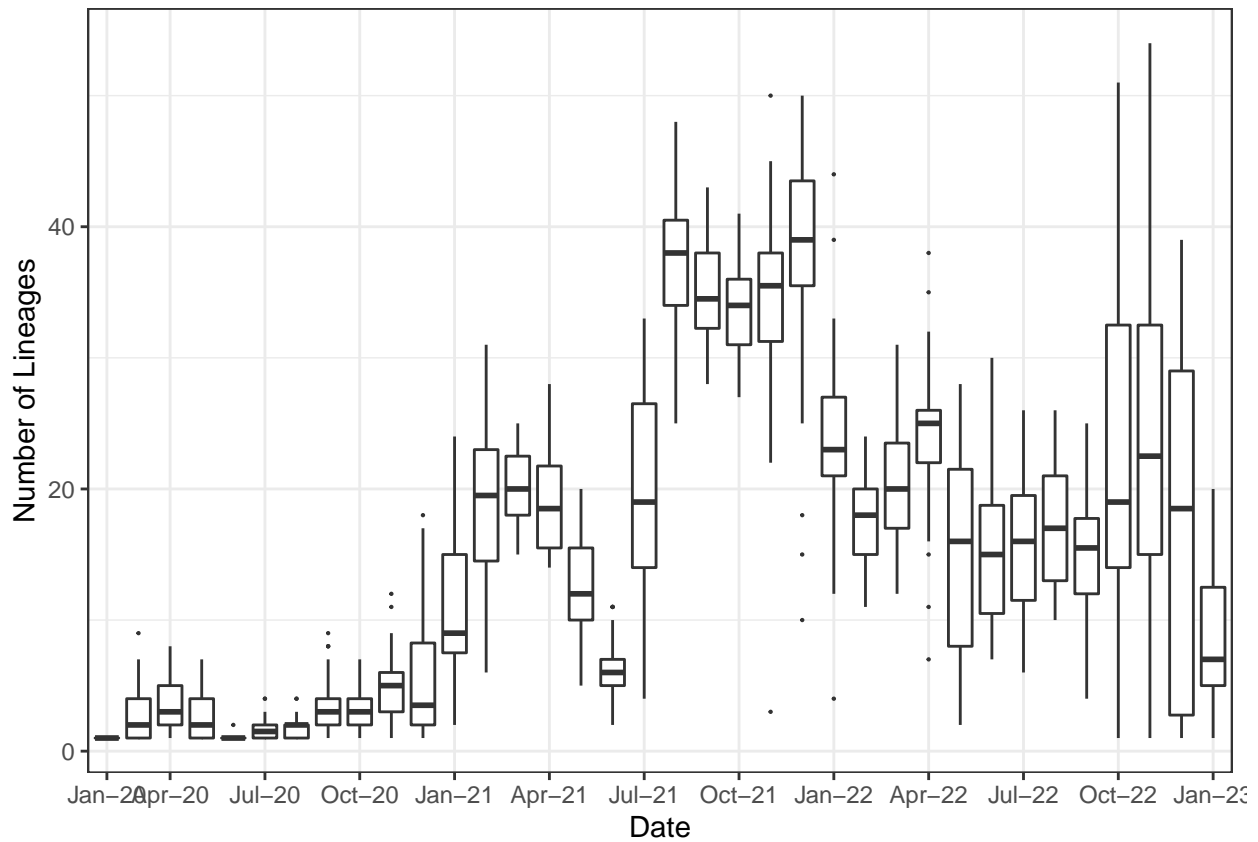
Lineage Proportions

Here we filtered the data to only include 5 lineages with the highest proportions in the samples taken on a given day for readability. We also truncated the dates to the period between March 2021 to Jan 2023.I decided to take out the legends since there are 315 unique lineages, the legends and the plot are still not very readable.
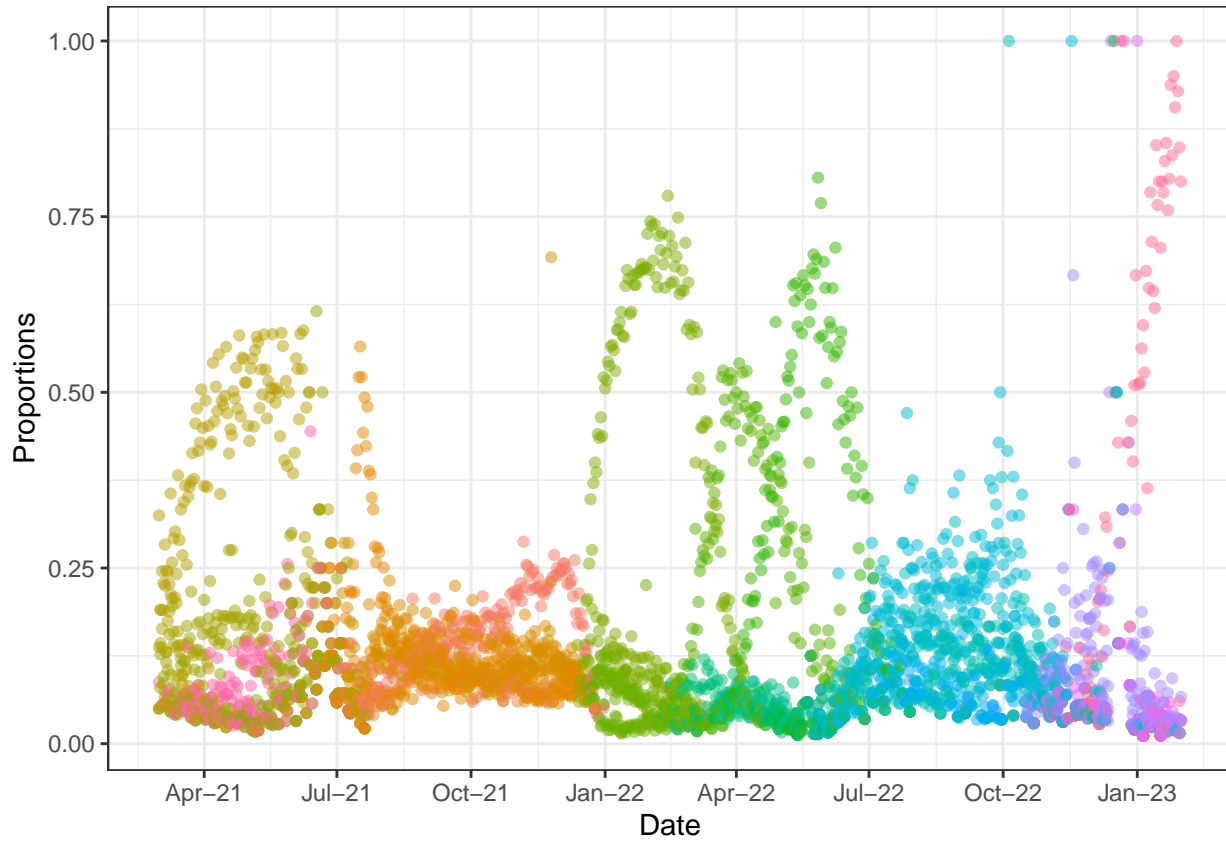
AY.116	B.1.1	BA.1.20	BA.2.75	BA.5.2	BF.31.1	BQ.1.13	CR.1
AY.117	B.1.1.316	BA.2	BA.2.75.2	BA.5.2.1	BF.4	BQ.1.13.1	CR.1.1
AY.118	B.1.1.318	BA.2.1	BA.2.75.5	BA.5.2.18	BF.5	BQ.1.14	DC.1
AY.119	B.1.1.519	BA.2.10	BA.2.76	BA.5.2.20	BF.7	BQ.1.15	DF.1
AY.119.2	B.1.1.7	BA.2.10.1	BA.2.81	BA.5.2.21	BF.7.4.1	BQ.1.2	DF.1.1
AY.120	B.1.2	BA.2.12	BA.2.9	BA.5.2.22	BF.7.6	BQ.1.23	P.1
AY.120.1	B.1.243	BA.2.12.1	BA.2.9.2	BA.5.2.3	BF.8	BQ.1.25	P.1.10
AY.122	B.1.351	BA.2.13	BA.2.9.3	BA.5.2.34	BG.5	BQ.1.27	P.1.13
AY.13	B.1.427	BA.2.13.1	BA.2.9.5	BA.5.2.35	BM.1	BQ.1.28	P.1.17
AY.14	B.1.429	BA.2.18	BA.4	BA.5.2.6	BN.1	BQ.1.3	P.1.4
AY.20	B.1.517	BA.2.2	BA.4.1	BA.5.2.9	BN.1.3	BQ.1.5	Q.4
AY.25	B.1.525	BA.2.20	BA.4.1.8	BA.5.3	BN.1.4	BQ.1.6	R.1
AY.25.1	B.1.526	BA.2.21	BA.4.1.9	BA.5.3.1	BN.1.9	BQ.1.8	XBB
AY.26	B.1.568	BA.2.22	BA.4.2	BA.5.3.5	BQ.1	BQ.1.9	XBB.1
AY.3	B.1.575	BA.2.23	BA.4.4	BA.5.5	BQ.1.1	BQ.2	XBB.1.5
AY.3.1	B.1.617.2	BA.2.23.1	BA.4.6	BA.5.5.1	BQ.1.1.1	BR.1	XBB.1.7
AY.3.3	B.1.621	BA.2.3	BA.4.6.2	BA.5.6	BQ.1.1.13	BU.1	XBB.1.9.1
AY.35	B.1.621.1	BA.2.3.17	BA.4.6.4	BA.5.6.1	BQ.1.1.18	C.36.3	XBB.2
AY.39	B.1.623	BA.2.3.20	BA.4.6.5	BE.1	BQ.1.1.19	C.36.3.1	XBB.3

Due to a lot of variability in the unique lineages being sampled daily, we need another way to make the plot more readable. We will only include 5 lineages with the highest proportions in the samples taken on a given epiweek (instead of day), so we can exclude lineages that are not consistently detected. With this, there are 88 unique lineages below
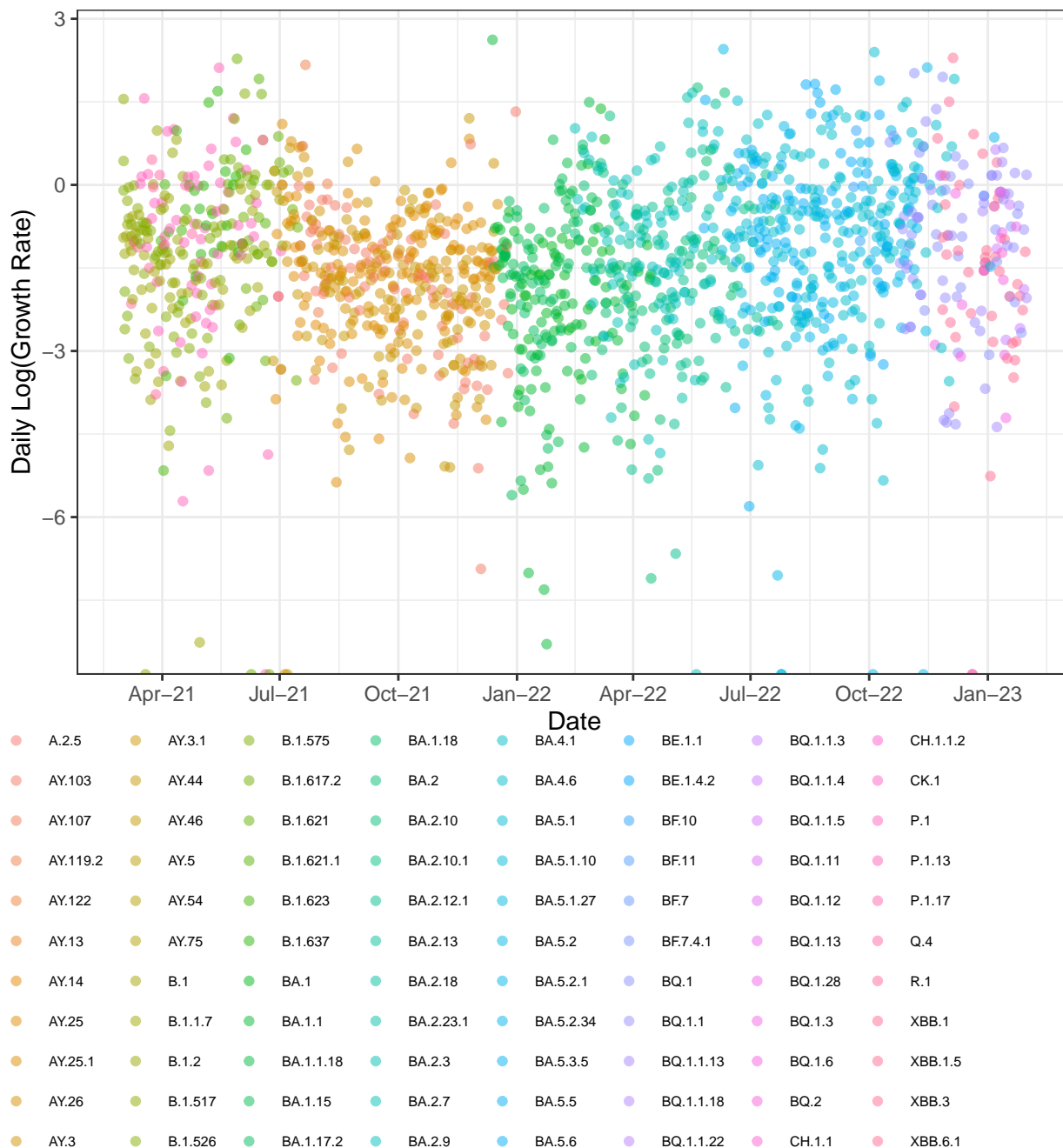
Sampling rates

We divided the total number of samples taken on a given day by the population of MA, which is a constant number of about 68 (multiplied by 100K). Again we will focus on the period starting March 2021 and after. The rates were fairly low from the later part of 2022 onward. The sampling rates will be used as a feature and probably can use some smoothing (smooth line in the figure generated from a cubic spline).
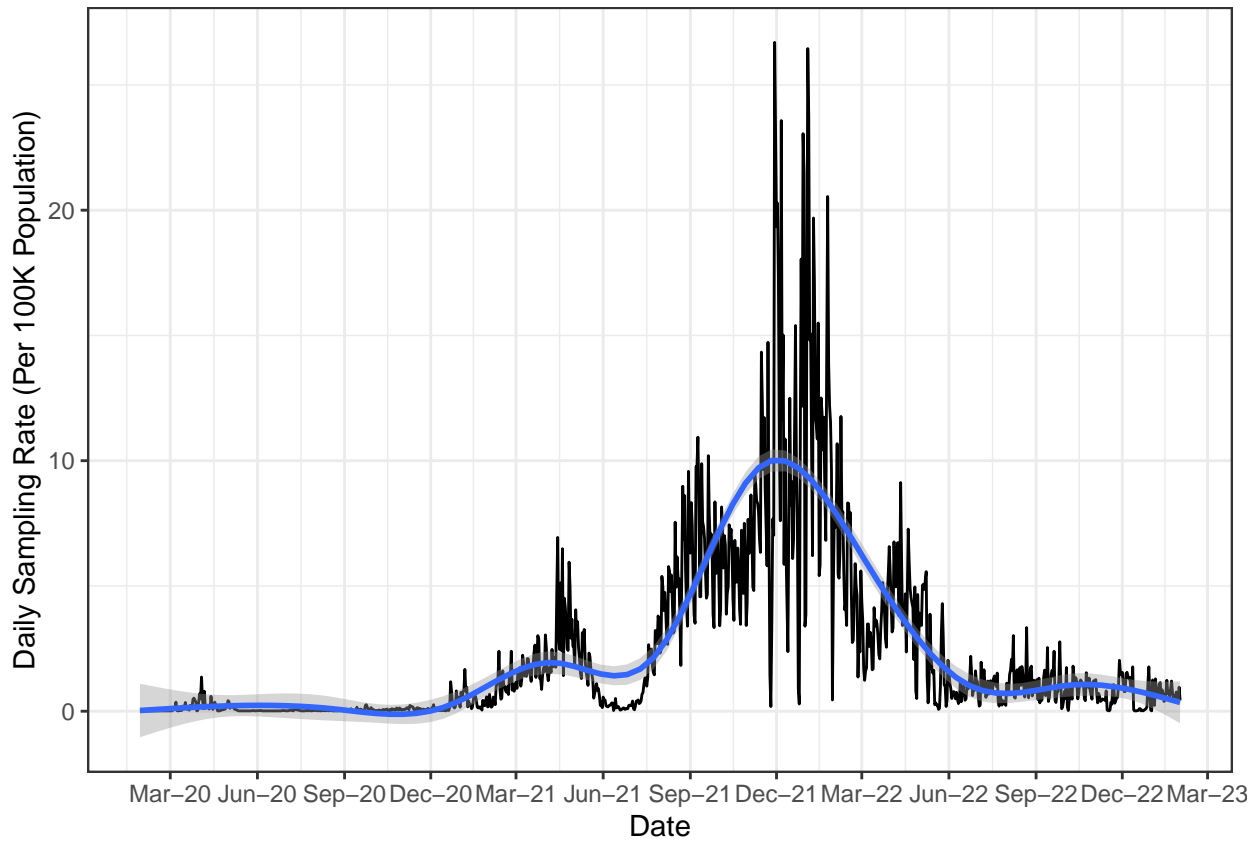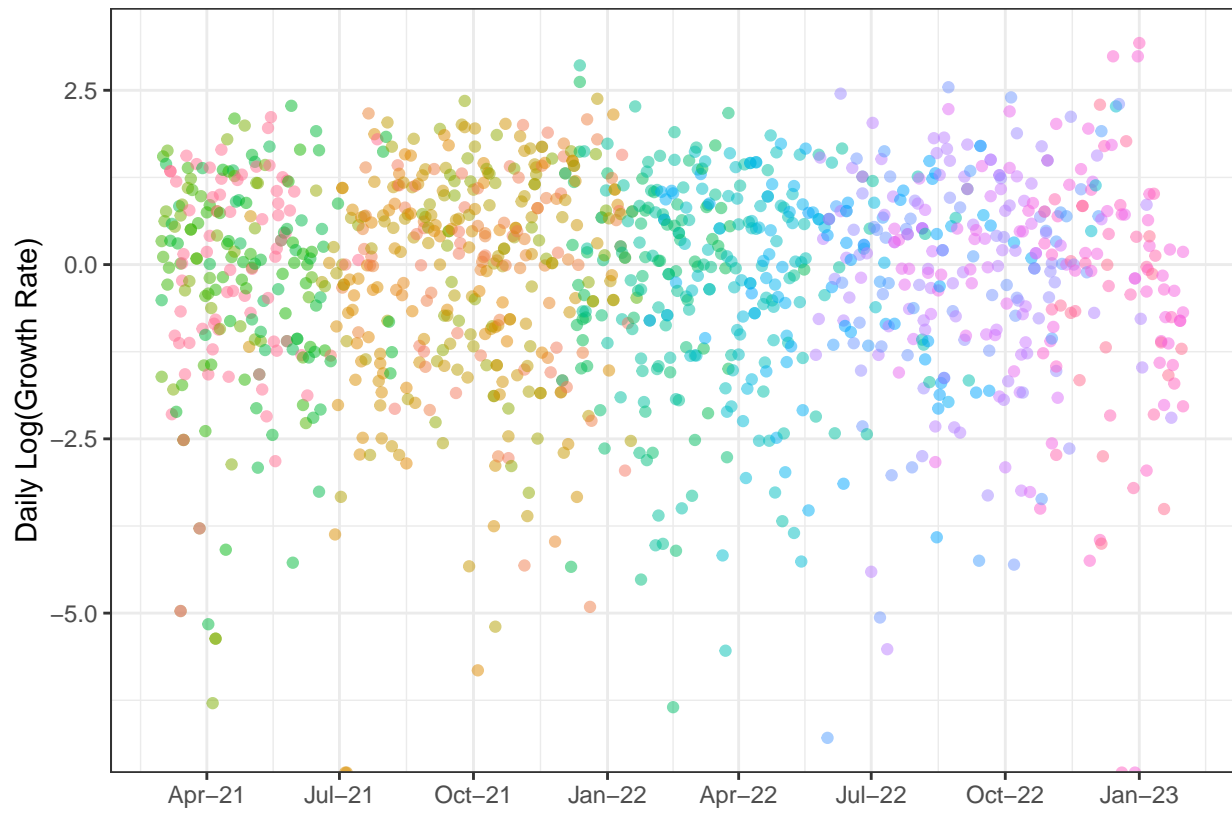
Growth Rates

By now I realized that the daily sampling rates fluctuate so much that the lineages detected varied greatly day to day. I decided to look at the daily growth rates of 5 lineages with the highest proportions in a given week (there are 88 unique lineages during this period), so we have relatively stable sets of lineages. Well, it's still not readable.

A.2.5	AY.3.1	B.1.575	BA.1.18	BA.4.1	BE.1.1	BQ.1.1.3	CH.1.1.2	
AY.103	AY.44	B.1.617.2	BA.2	BA.4.6	BE.1.4.2	BQ.1.1.4	CK.1	
AY.107	AY.46	B.1.621	BA.2.10	BA.5.1	BF.10	BQ.1.1.5	P.1	
AY.119.2	AY.5	B.1.621.1	BA.2.10.1	BA.5.1.10	BF.11	BQ.1.11	P.1.13	
AY.122	AY.54	B.1.623	BA.2.12.1	BA.5.1.27	BF.7	BQ.1.12	P.1.17	
AY.13	AY.75	B.1.637	BA.2.13	BA.5.2	BF.7.4.1	BQ.1.13	Q.4	
AY.14	B.1	BA.1	BA.2.18	BA.5.2.1	BQ.1	BQ.1.28	R.1	
AY.25	B.1.1.7	BA.1.1	BA.2.23.1	BA.5.2.34	BQ.1.1	BQ.1.3	XBB.1	
AY.25.1	B.1.2	BA.1.1.18	BA.2.3	BA.5.3.5	BQ.1.1.13	BQ.1.6	XBB.1.5	
AY.26	B.1.517	BA.1.15	BA.2.7	BA.5.5	BQ.1.1.18	BQ.2	XBB.3	
AY.3	B.1.526	BA.1.17.2	BA.2.9	BA.5.6	BQ.1.1.22	CH.1.1	XBB.6.1	

And I did the same but with 5 lineages with highest average daily growth rates in a given week. There are 209 unique lineages, too many to be readable on the plot below.