Table of Contents

[2021/09/10 Data 2](#_Toc98884059)

[Methods 2](#_Toc98884060)

[Data Acquisition 2](#_Toc98884061)

[Download from NextStrain 2](#_Toc98884062)

[Extract the GISAID accession IDs of the NextStrain subsamples 2](#_Toc98884063)

[Download NextStrain subsample sequence from GISAID 2](#_Toc98884064)

[Multiple sequence alignment and mutation calling via NextClade 2](#_Toc98884065)

[Other… 2](#_Toc98884066)

[Data Analysis 3](#_Toc98884067)

[Sample Characteristic 3](#_Toc98884068)

[Clade Identification 3](#_Toc98884069)

[Clade Parameters Calculation 3](#_Toc98884070)

[Survival Analysis Sample Acquisition 3](#_Toc98884071)

[Results 5](#_Toc98884072)

[All sequence samples 5](#_Toc98884073)

[Time span 5](#_Toc98884074)

[Phylogenetic tree 5](#_Toc98884075)

[Frequency of substitutions in tips (samples) 5](#_Toc98884076)

[E484 5](#_Toc98884077)

[N501Y 5](#_Toc98884078)

[Frequency of substitutions in ancestral nodes (ASR) 5](#_Toc98884079)

[E484 5](#_Toc98884080)

[N501 5](#_Toc98884081)

[E484-mutate clades characteristics – raw 5](#_Toc98884082)

[Size 5](#_Toc98884083)

[Duration 6](#_Toc98884084)

[Time distribution of samples in each clade 6](#_Toc98884085)

[Time distribution of clade emergence (by earliest sample) 6](#_Toc98884086)

[Clade duration and size vs. time of emergence (by earliest sample) 7](#_Toc98884087)

[E484-mutate clades genetic background 8](#_Toc98884088)

[Proportion of potential epistatic mutations among all E484K sequences (≥ 10%) 8](#_Toc98884089)

[Proportion of potential epistatic mutations in each E484K clade (≥ 10%) 8](#_Toc98884090)

[E484-mutate clade emergence association with time (epidemiology: level of immunity) 9](#_Toc98884091)

[Standardized emergence by time (unit = 2-week) 9](#_Toc98884092)

[Standardized emergence by time - 1006 Data 9](#_Toc98884093)

[Logistic regression by continuous time (unit = day) 10](#_Toc98884094)

[Logistic regression by continuous time (unit = week) 10](#_Toc98884095)

[Logistic regression by continuous time (unit = month) 10](#_Toc98884096)

[E484-mutate clade emergence association with N501Y (epistasis) 11](#_Toc98884097)

[Proportion of N501 state at ancestral nodes of all E484 clades 11](#_Toc98884098)

[Logistic regression of E484K emergence by N501 state at ancestral node 12](#_Toc98884099)

[Survival analysis of E484K emergence by N501 state at parent of ancestral node 12](#_Toc98884100)

[Survival analysis of N501Y emergence by E484 state at parent of ancestral node 12](#_Toc98884101)

# **2021/09/10 Data**

## **Methods**

### Data Acquisition

Folder: /Users/blahblah/Now/Harvard/E484K-20210910

#### Download from NextStrain

LINK: https://nextstrain.org/ncov/gisaid/global

> Go to bottom > DOWNLOAD DATA

> TIMETREE (NEWICK), ACKNOWLEDGEMENTS (TSV),

SCREENSHOT (SVG)

Outputs: subsample tree (nextstrain\_timetree\_210910.nwk)

sequence source information (nextstrain\_210910.tsv)

Date: 09/10/2021

#### Extract the GISAID accession IDs of the NextStrain subsamples

Command line: cut -f 2 nextstrain\_210910.tsv > nextstrain\_id\_210910.tsv

Outputs: nextstrain\_id\_210910.tsv

#### Download NextStrain subsample sequence from GISAID

LINK: https://www.gisaid.org/

> Search > Select > Upload nextstrain\_id\_210910.csv > OK > Download

Outputs: raw sequence (nextstrain\_210910.fasta)

time & location information (nextstrain\_time\_location\_210910.tsv)

patient metadata (nextstrain\_patient\_meta\_210910.tsv)

Date: 03/20/2022 (03/20/2022 metadata)

#### Multiple sequence alignment and mutation calling via NextClade

LINK: https://clades.nextstrain.org/

Select Pathogen: SARS-CoV-2 > Upload nextstrain\_210910.fasta

Outputs: aligned sequences (nextstrain\_nextclade\_aligned\_210910.fasta)

mutation calling (nextclade\_210910.csv)

Date: 03/20/2022

#### Other…

### Data Analysis

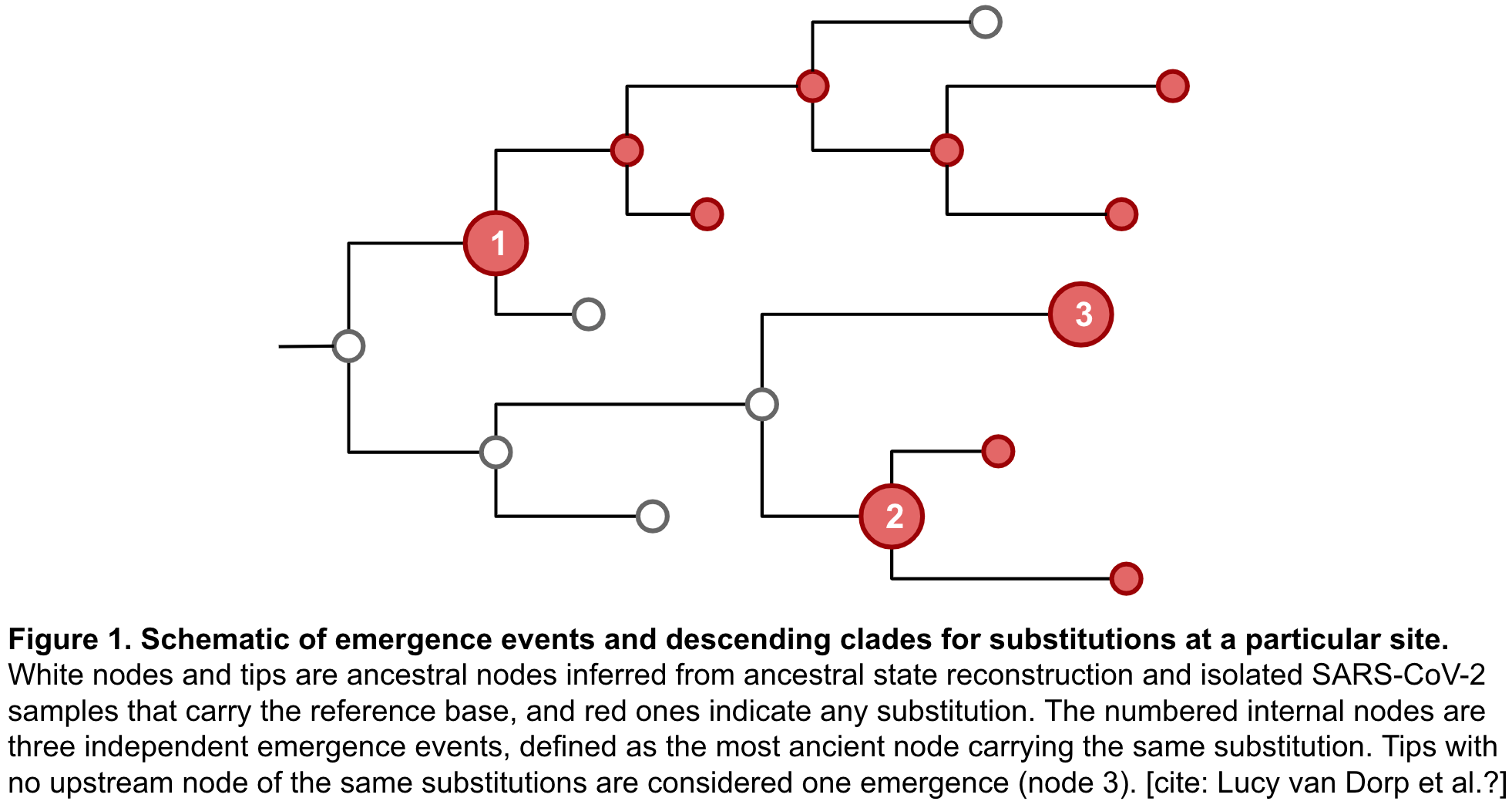
All data analysis performed using R version 4.0.2 (2020-06-22)

#### Sample Characteristic

Phylogenetic tree built using “ggtree” package

#### Clade Identification

Ancestral state reconstruction built using “castor” package



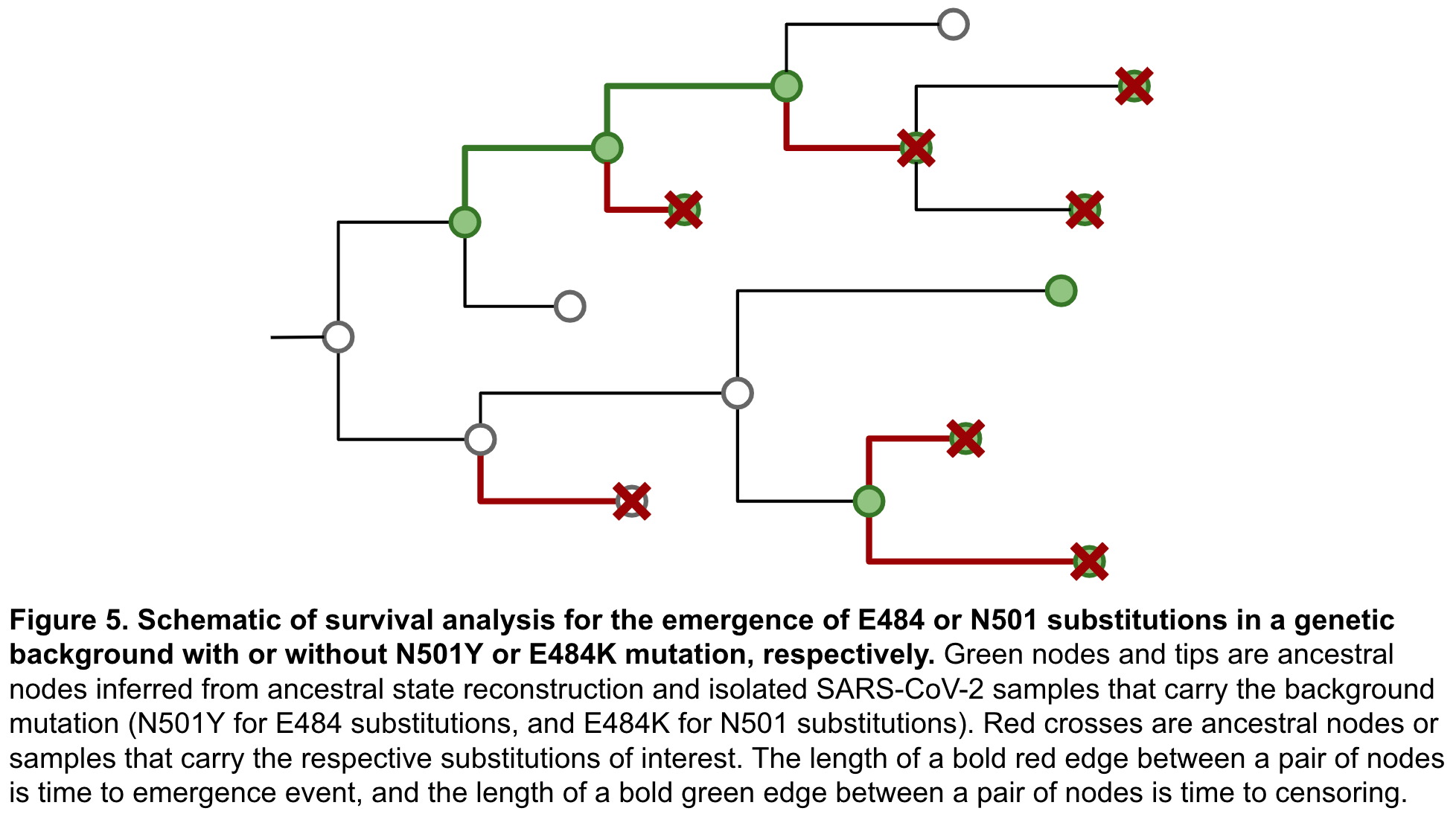
#### Clade Parameters Calculation

#### Survival Analysis Sample Acquisition

Survival analysis performed using “survival” package

**QUESTION: if interested in E484K emergence following N501Y (parent of ancestral**

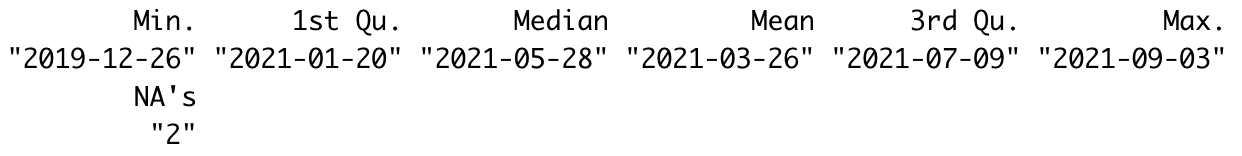
**node), neglect N501 status at the ancestral node?**



## **Results**

### All sequence samples

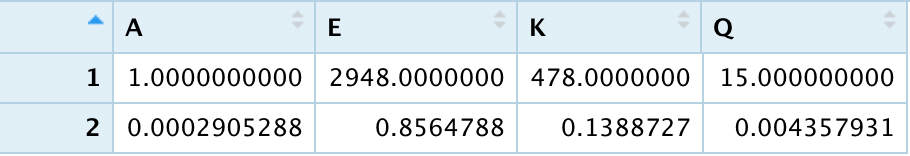
#### Time span

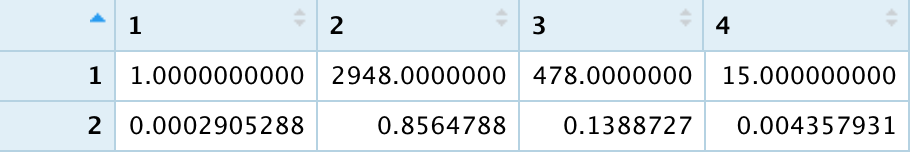


#### Phylogenetic tree

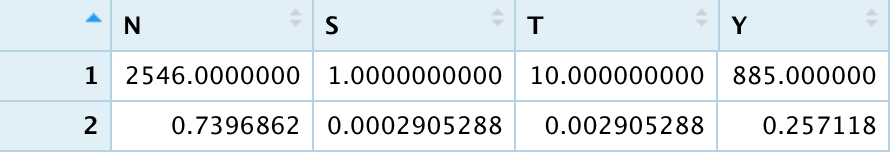
### Frequency of substitutions in tips (samples)

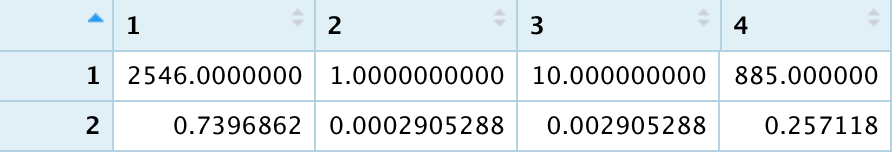
#### E484





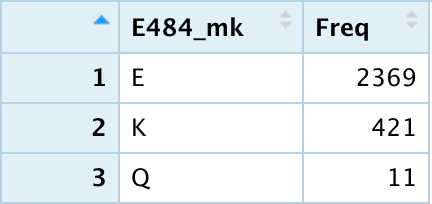
#### N501Y



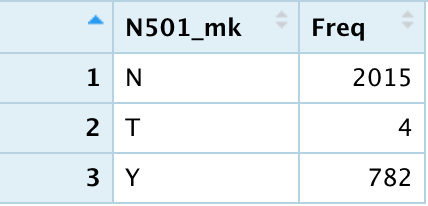


### Frequency of substitutions in ancestral nodes (ASR)

#### E484

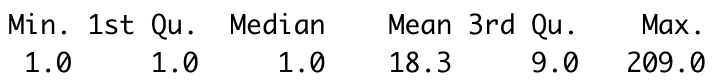


#### N501

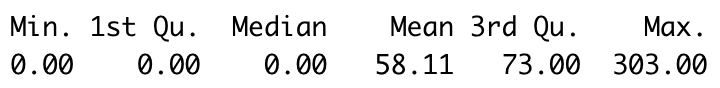


### E484-mutate clades characteristics – raw

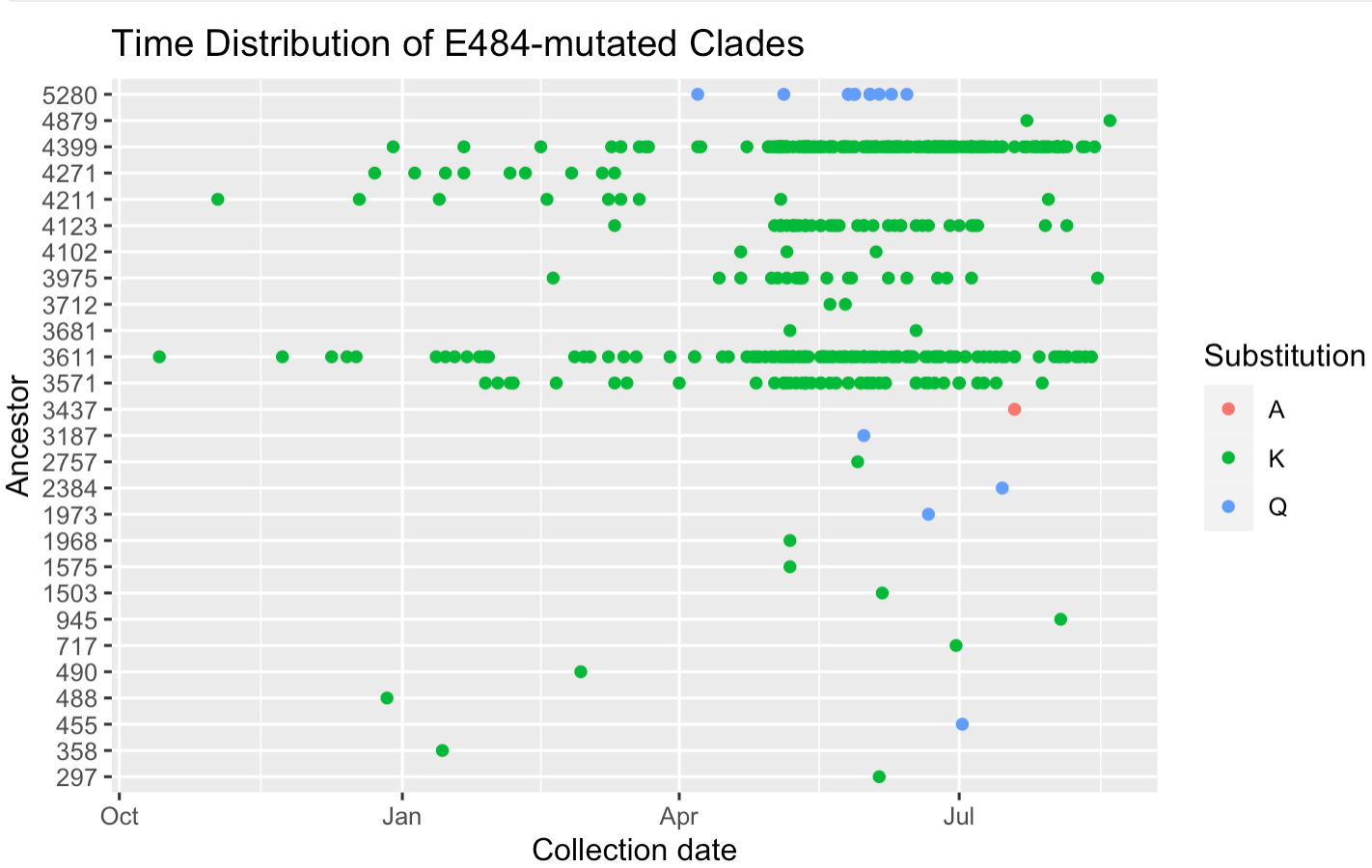
#### Size



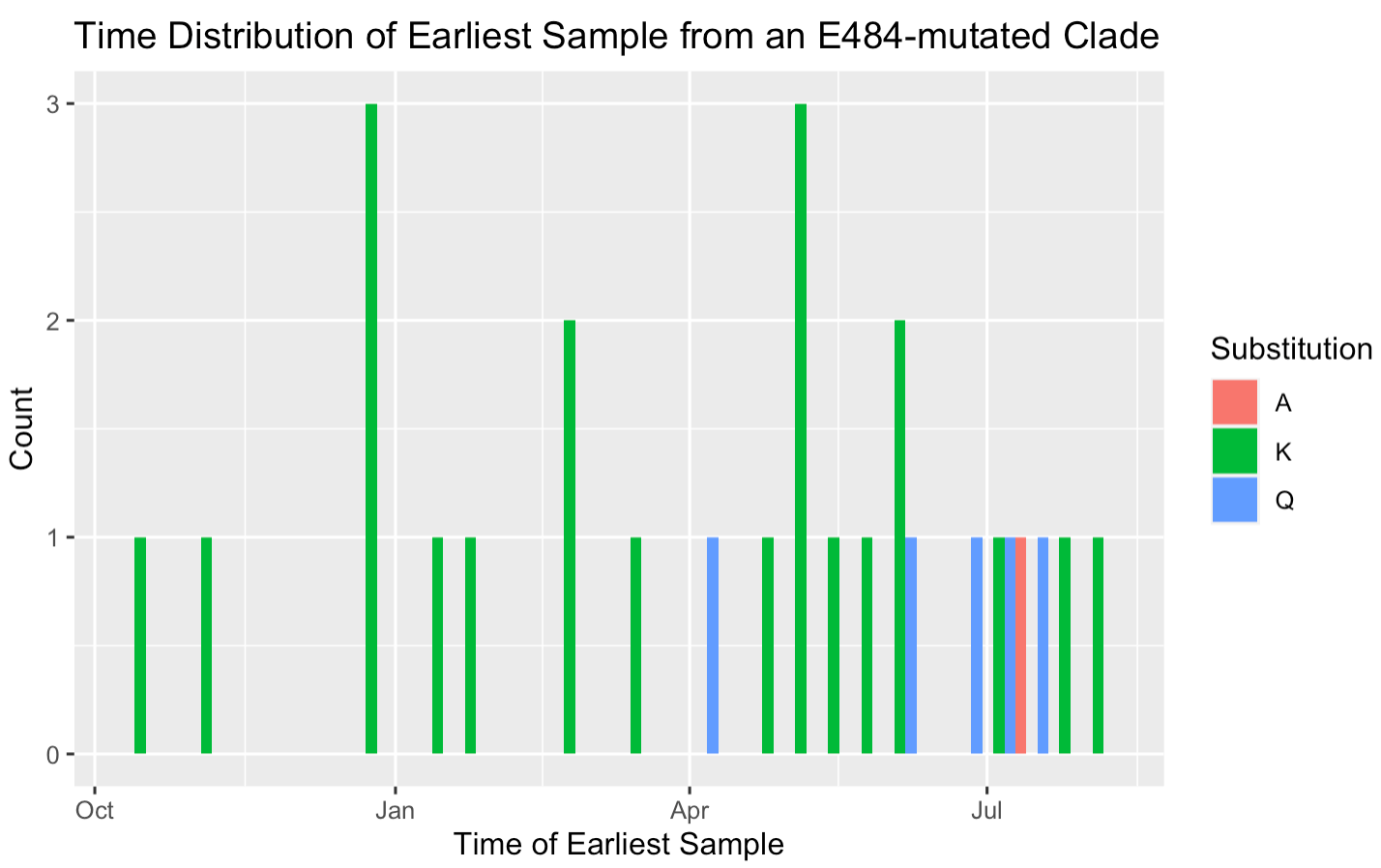
#### Duration

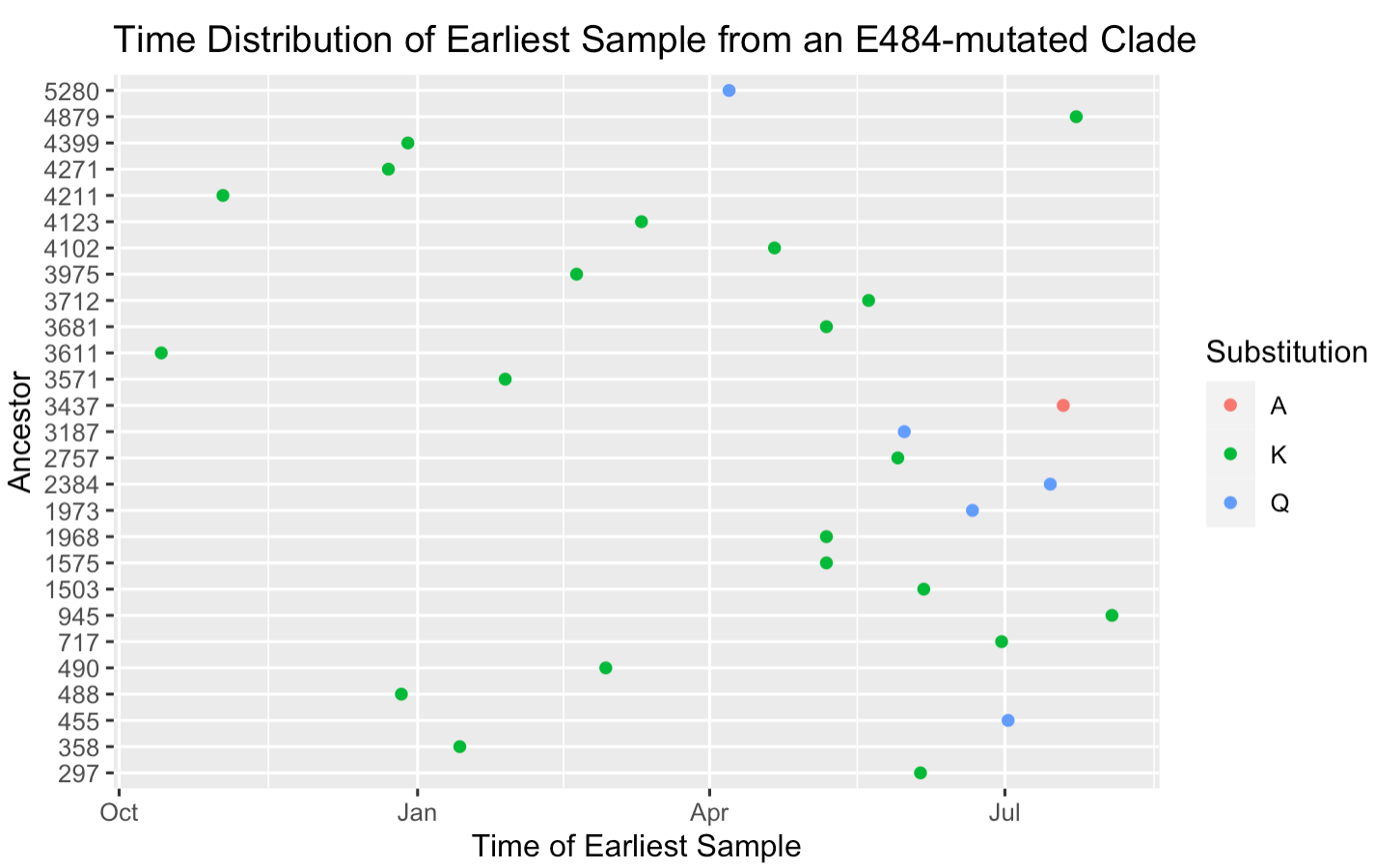


#### Time distribution of samples in each clade

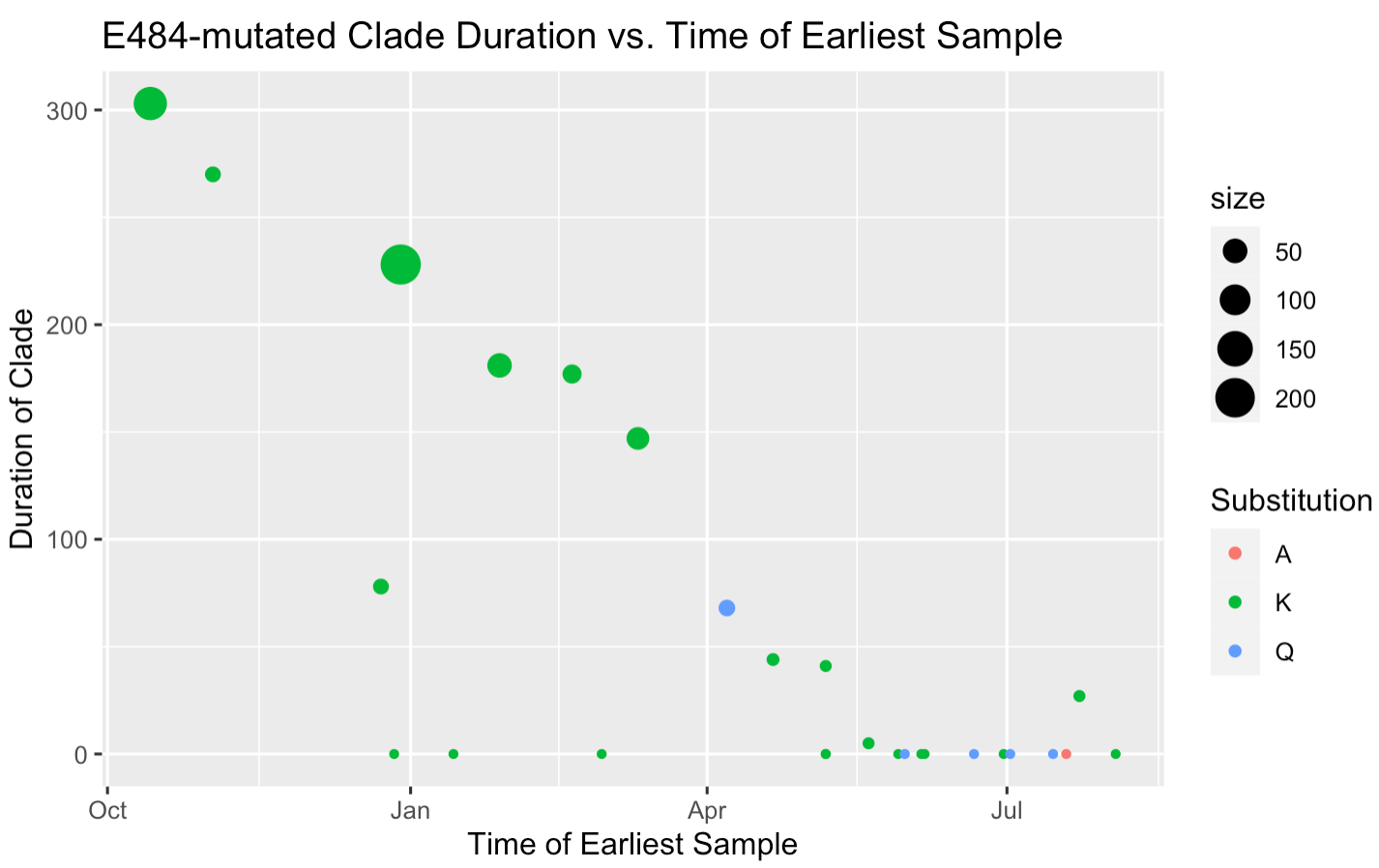


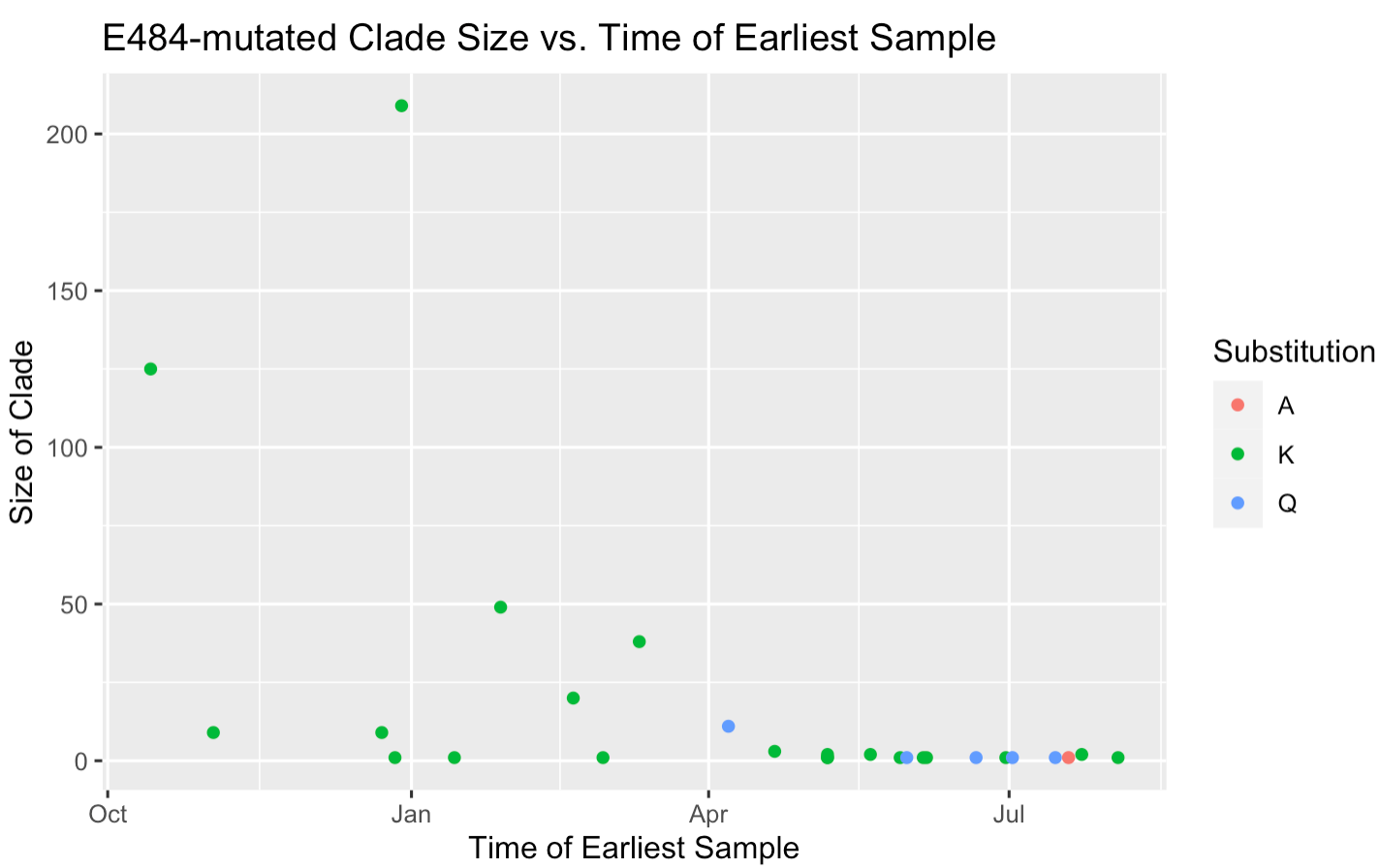
#### Time distribution of clade emergence (by earliest sample)





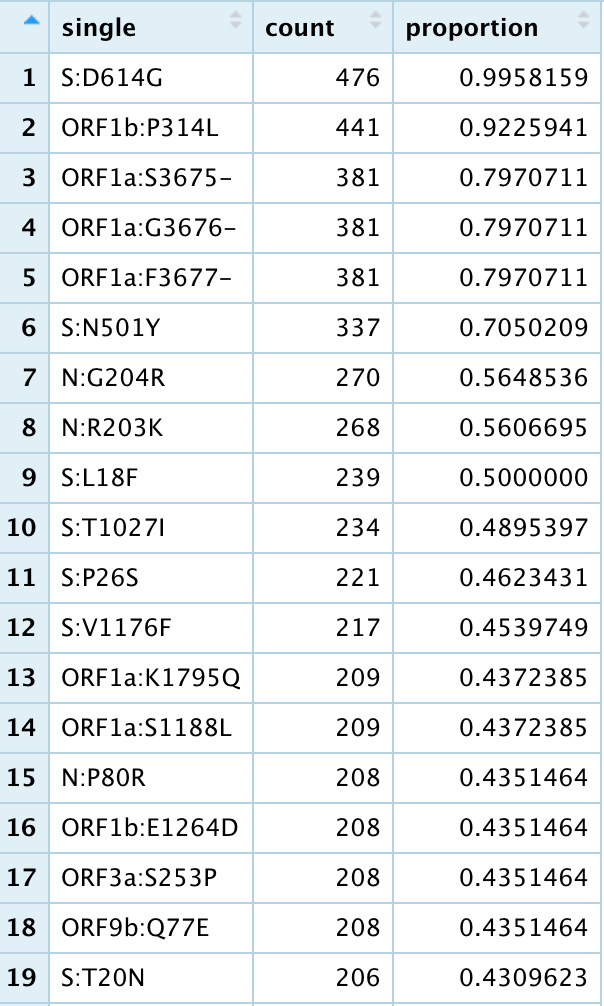
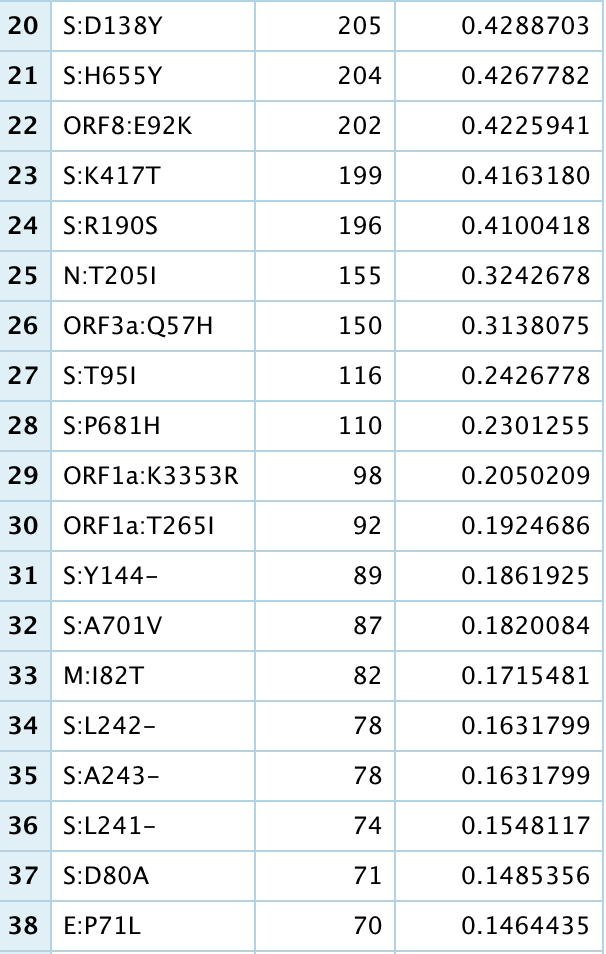
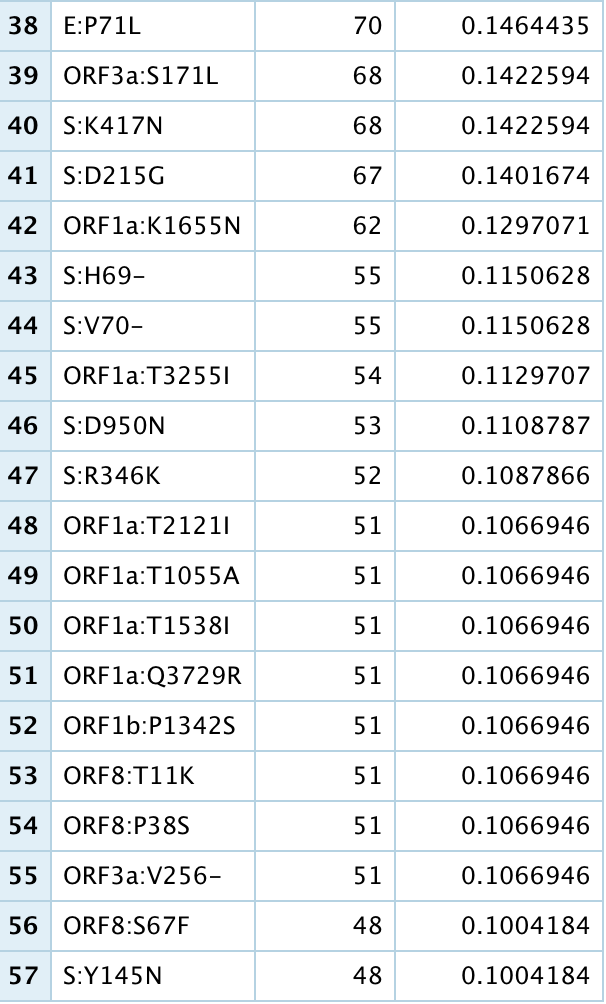
#### Clade duration and size vs. time of emergence (by earliest sample)



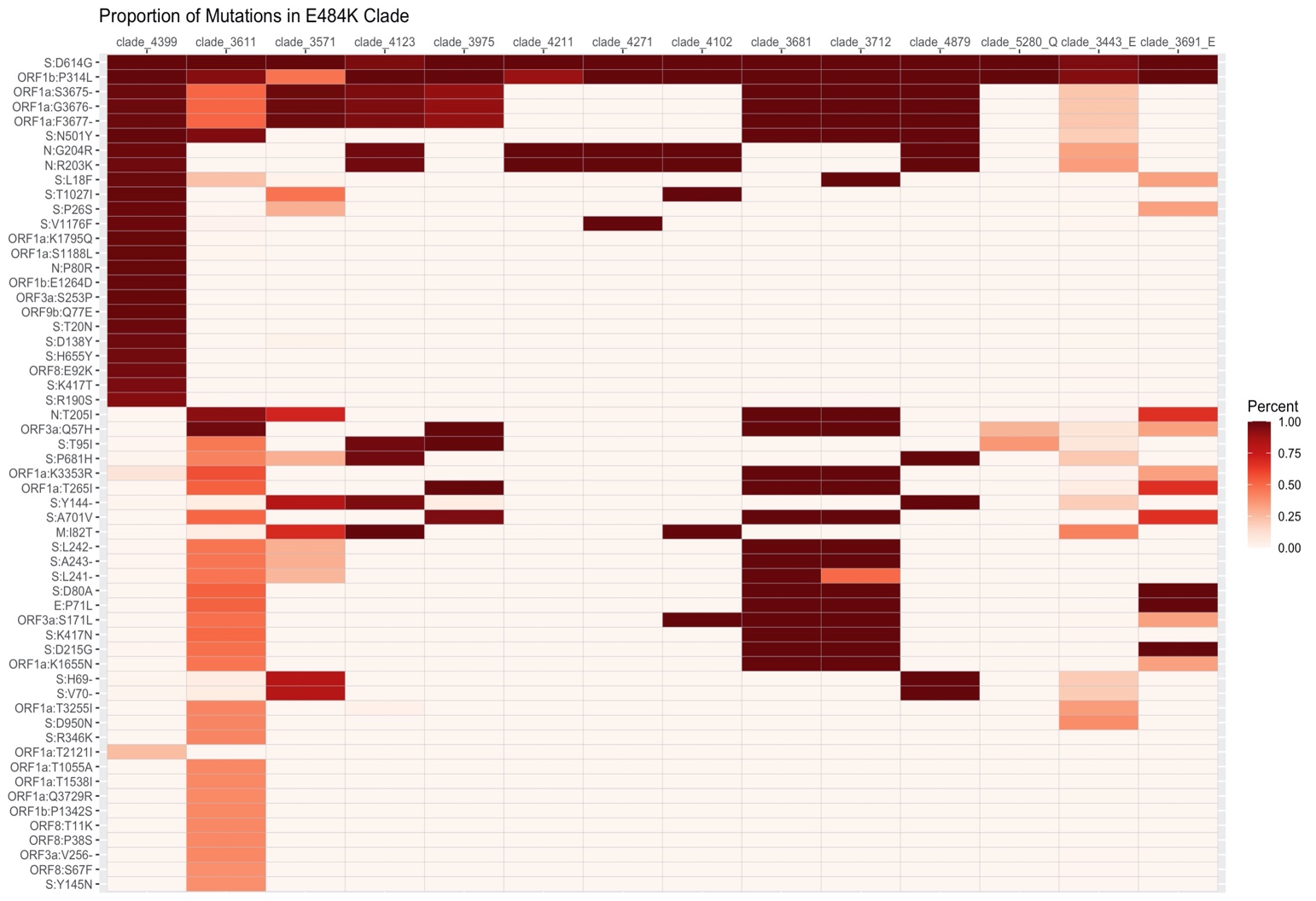


### E484-mutate clades genetic background

#### Proportion of potential epistatic mutations among all E484K sequences (≥ 10%)

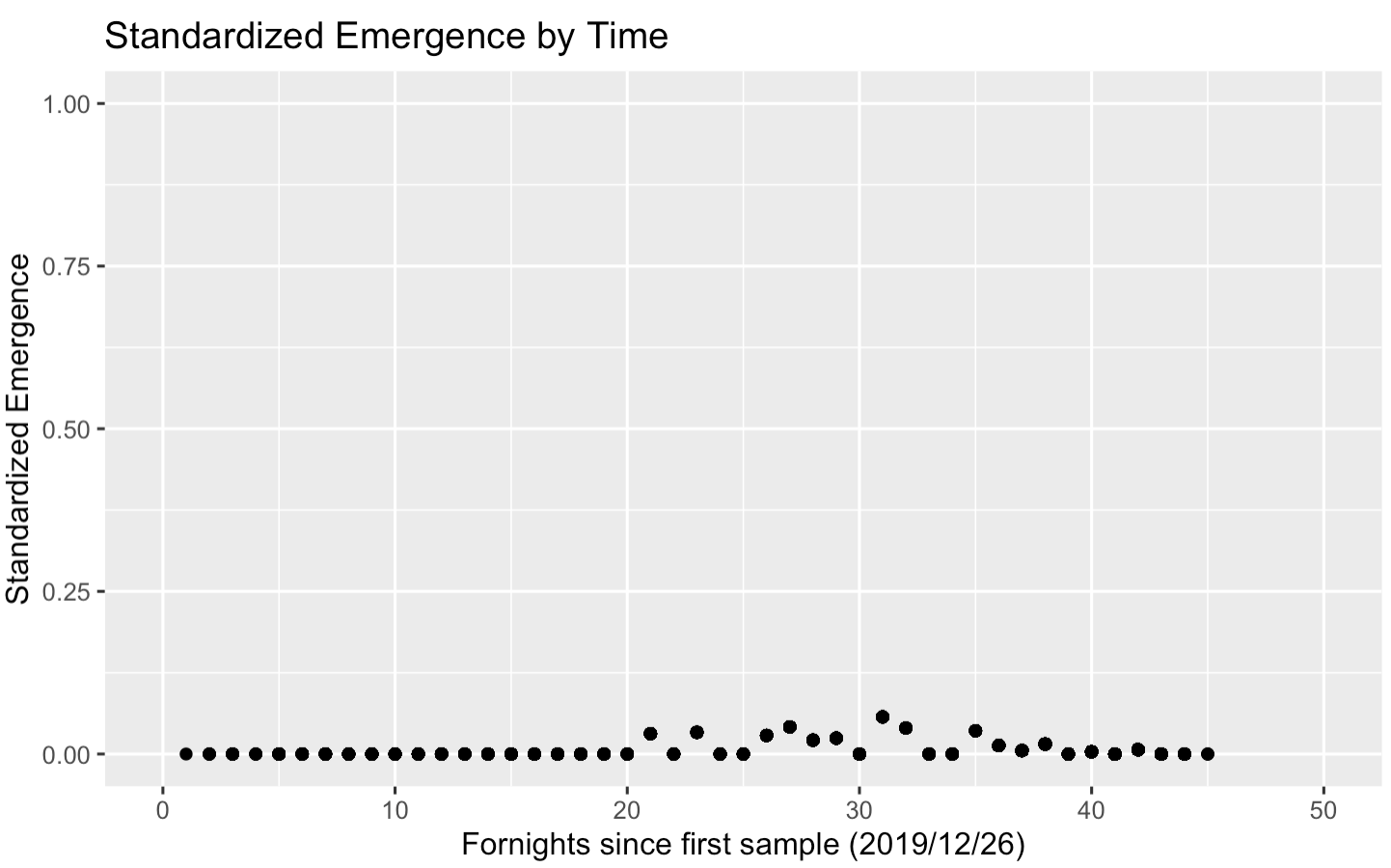
  

#### Proportion of potential epistatic mutations in each E484K clade (≥ 10%)



### E484-mutate clade emergence association with time (epidemiology: level of immunity)

#### Standardized emergence by time (unit = 2-week)

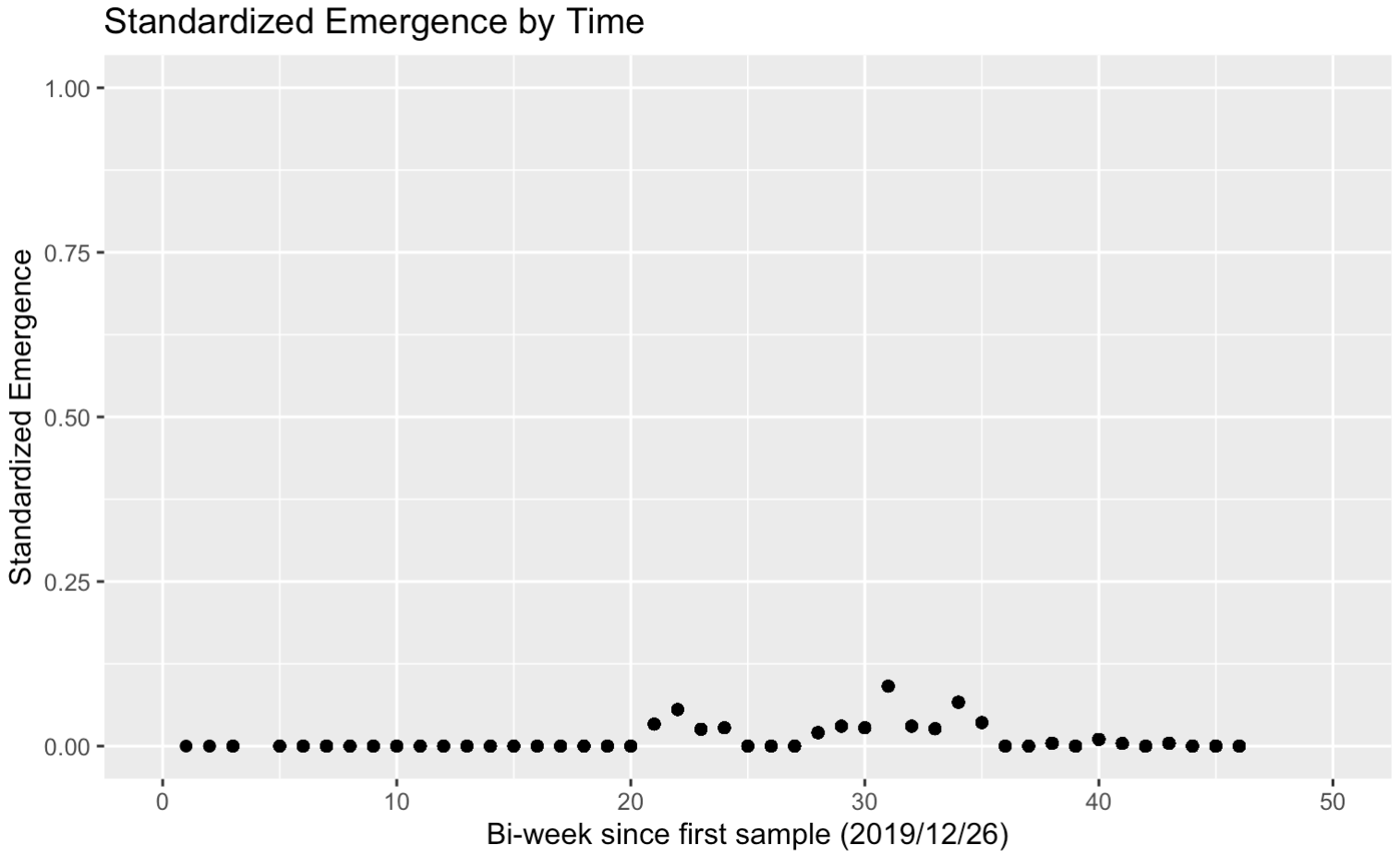


2020-09-21 ~ 2020-12-10

2020-12-10 ~ 2021-02-18

2021-02-18 ~ 2021-04-28

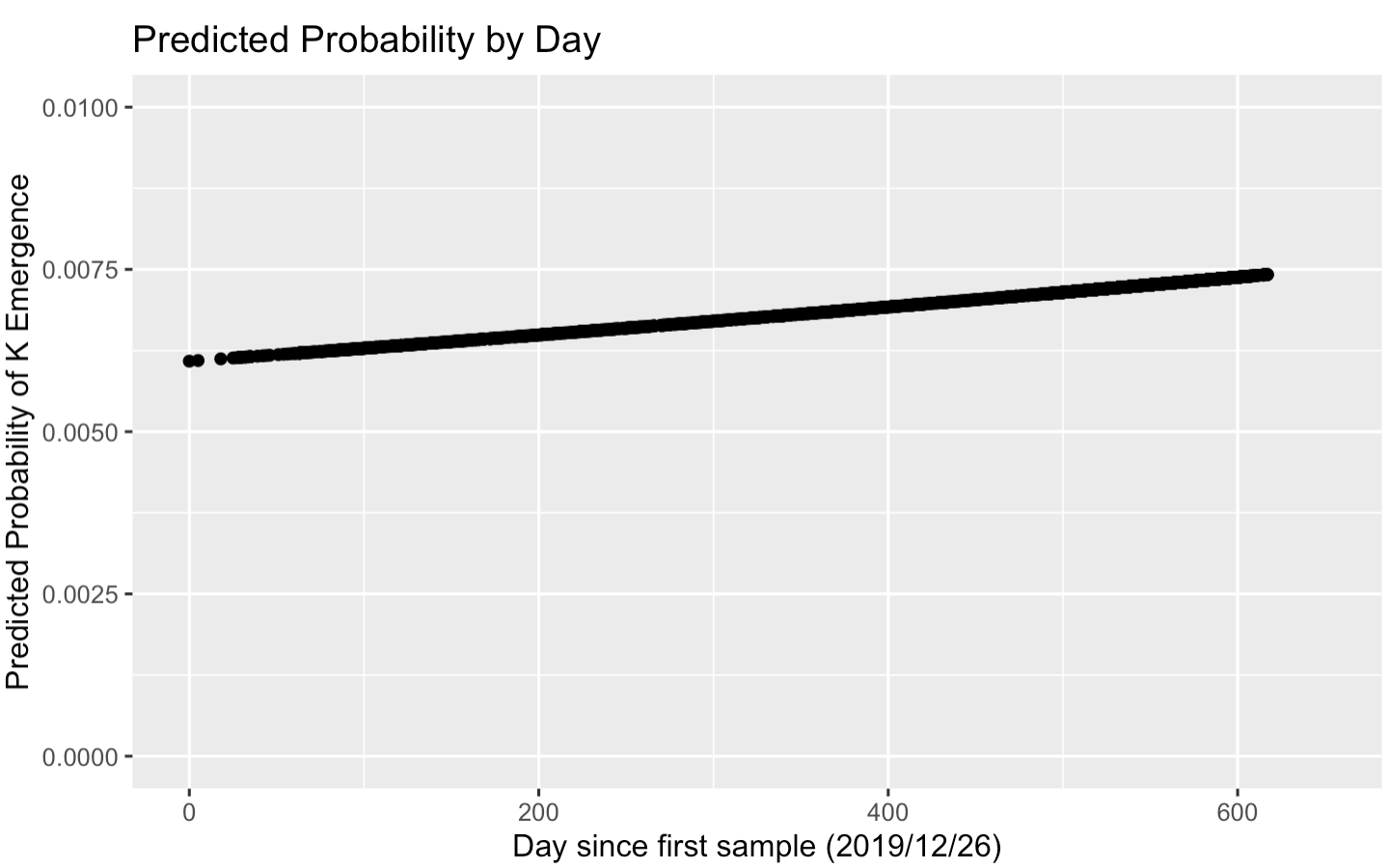
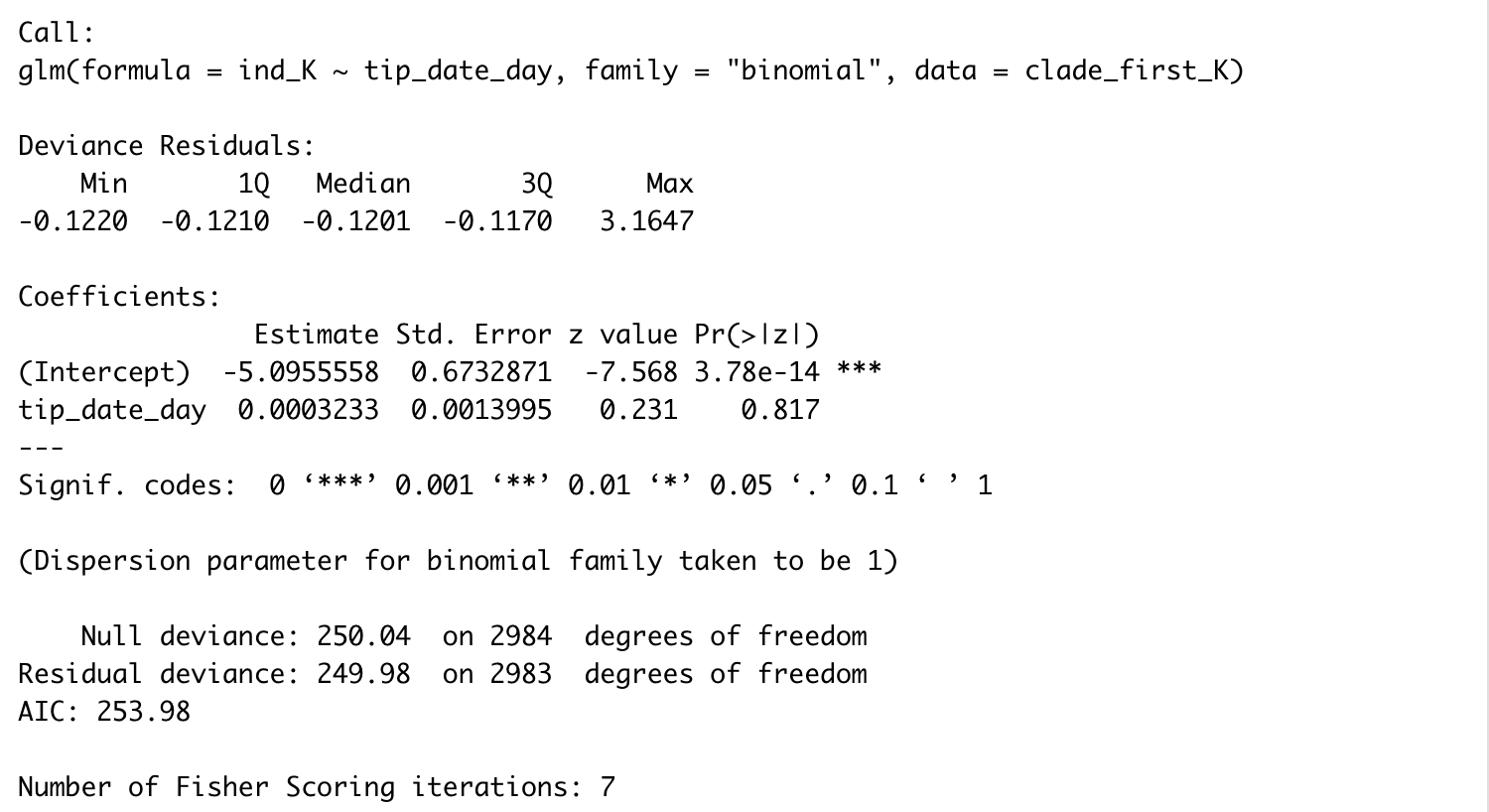
#### Standardized emergence by time - 1006 Data



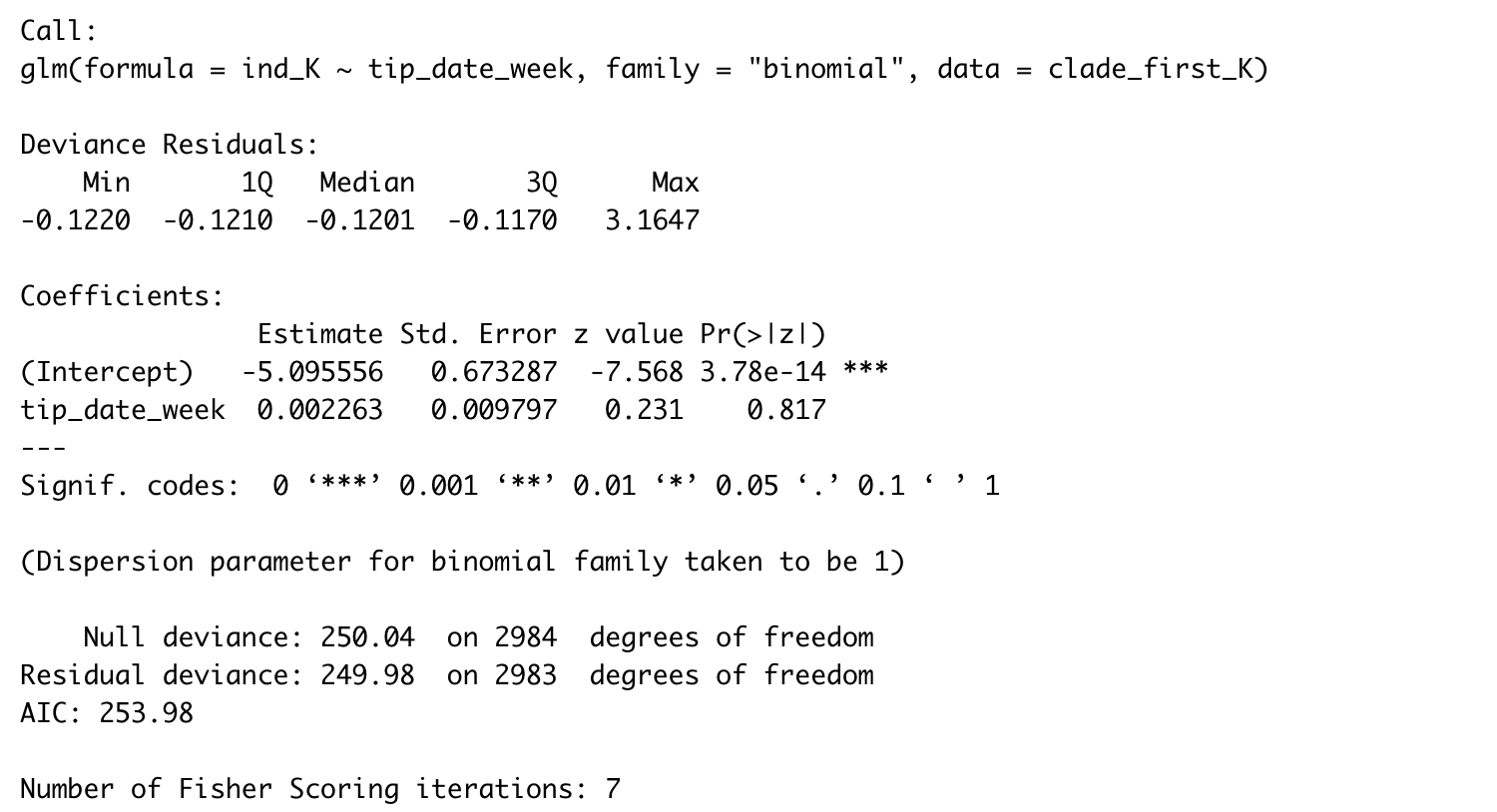
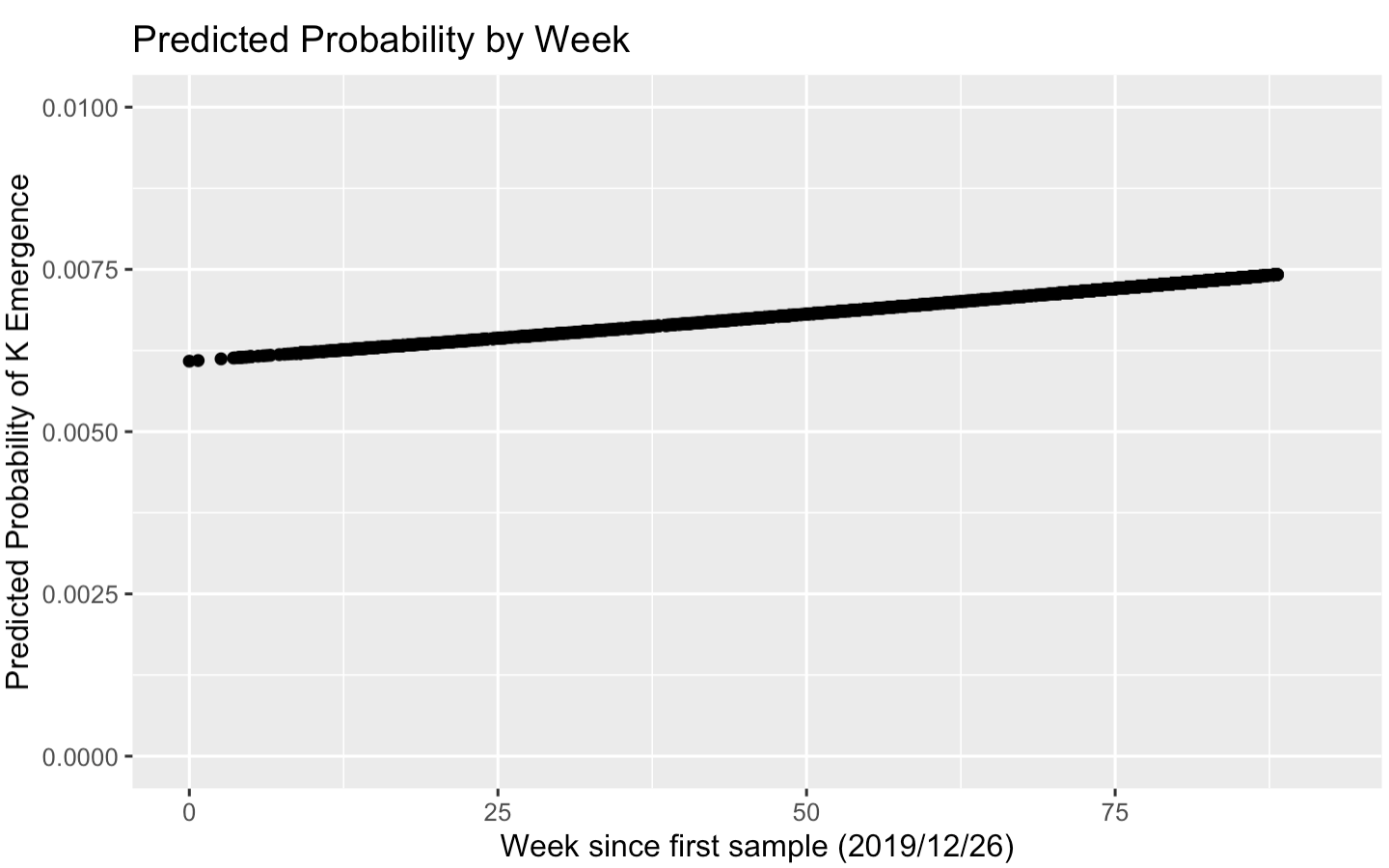
2020-09-18 ~ 2020-12-10

2020-12-10 ~ 2021-04-29

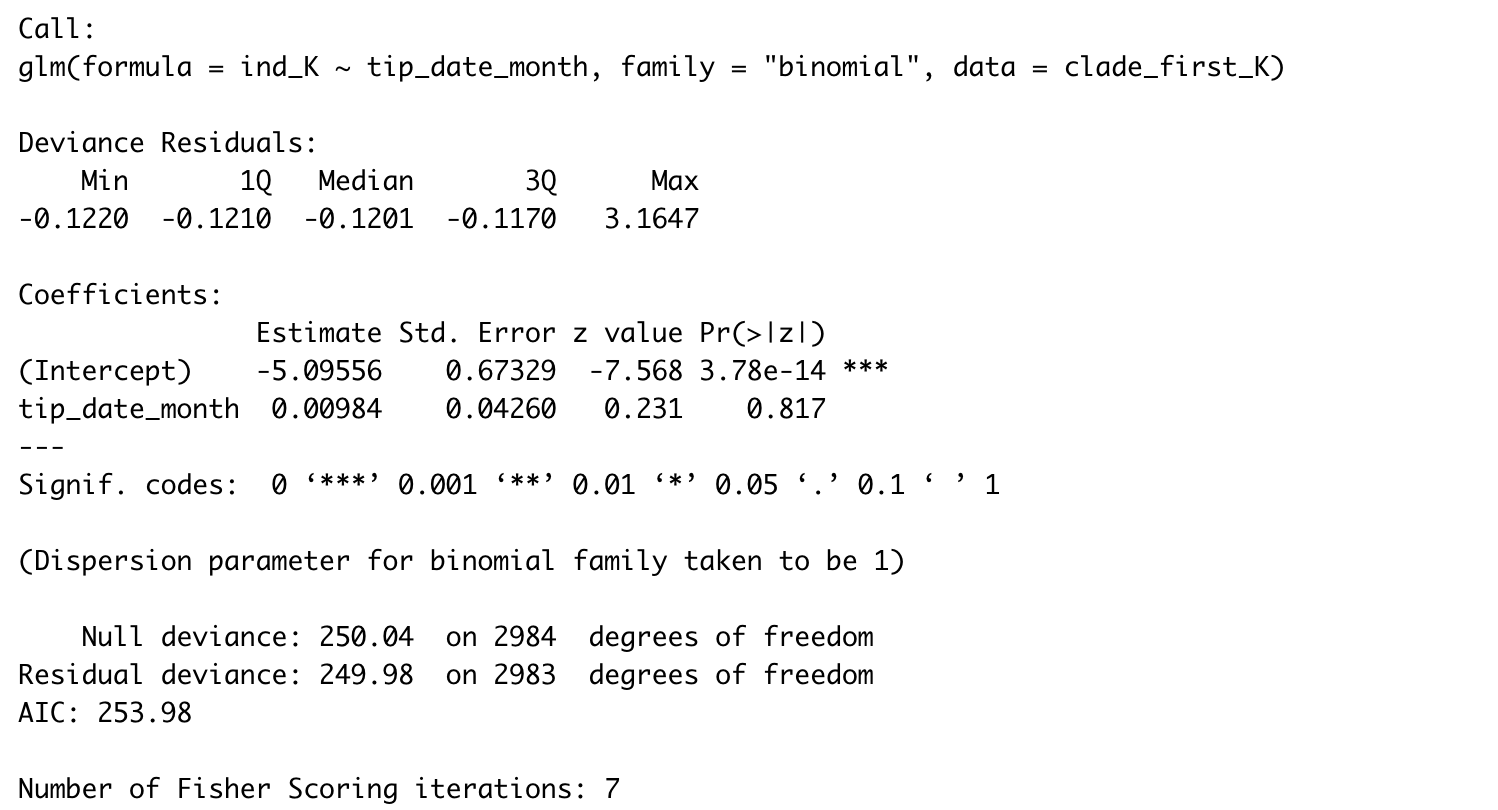
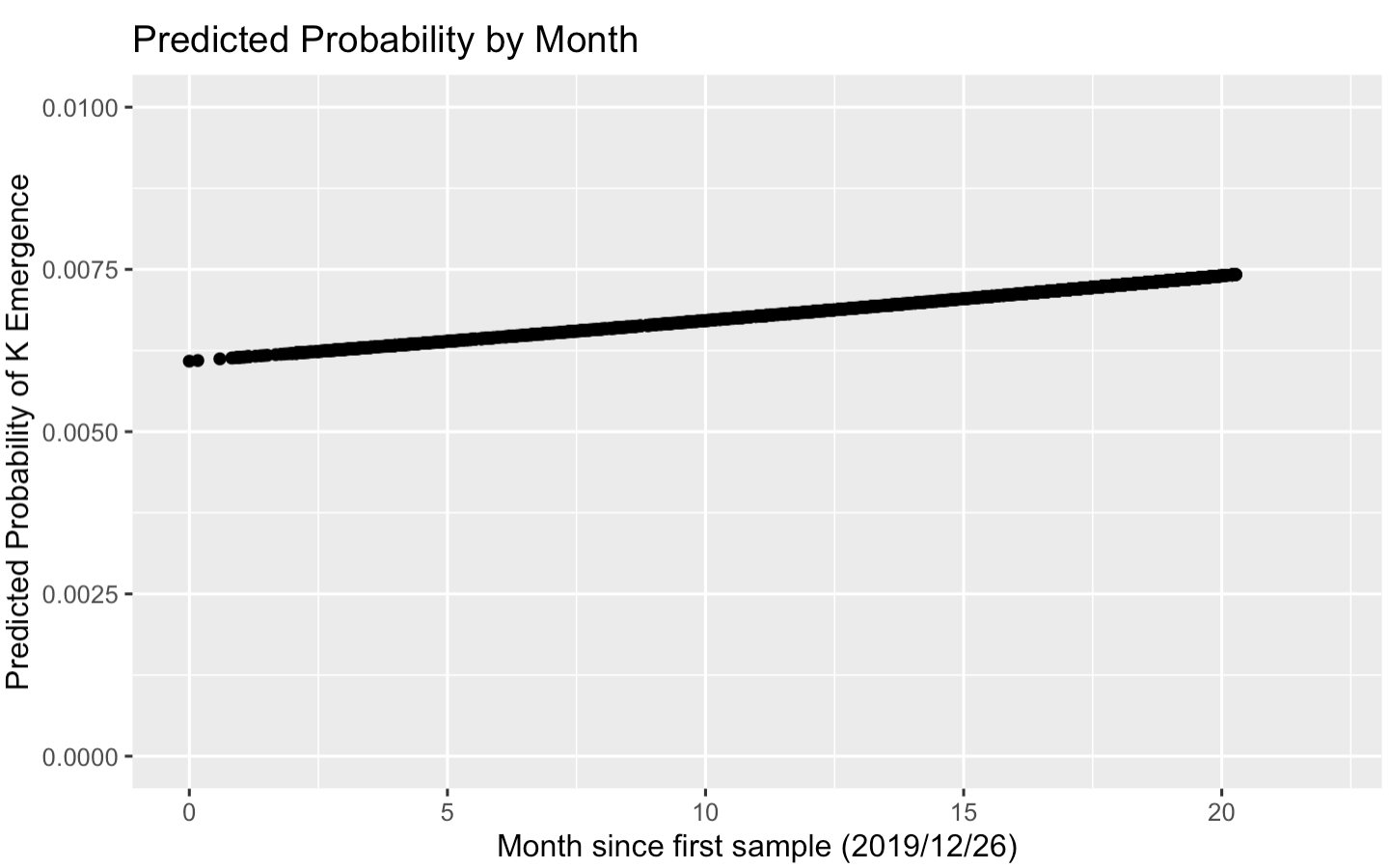
#### Logistic regression by continuous time (unit = day)



#### Logistic regression by continuous time (unit = week)

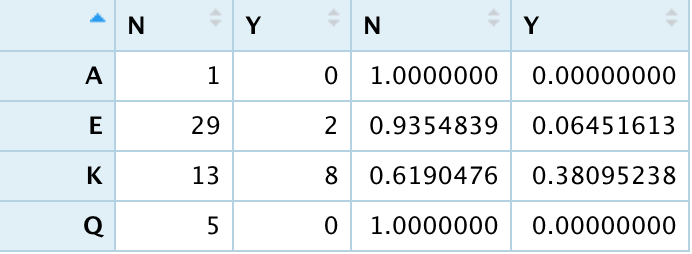
 

#### Logistic regression by continuous time (unit = month)

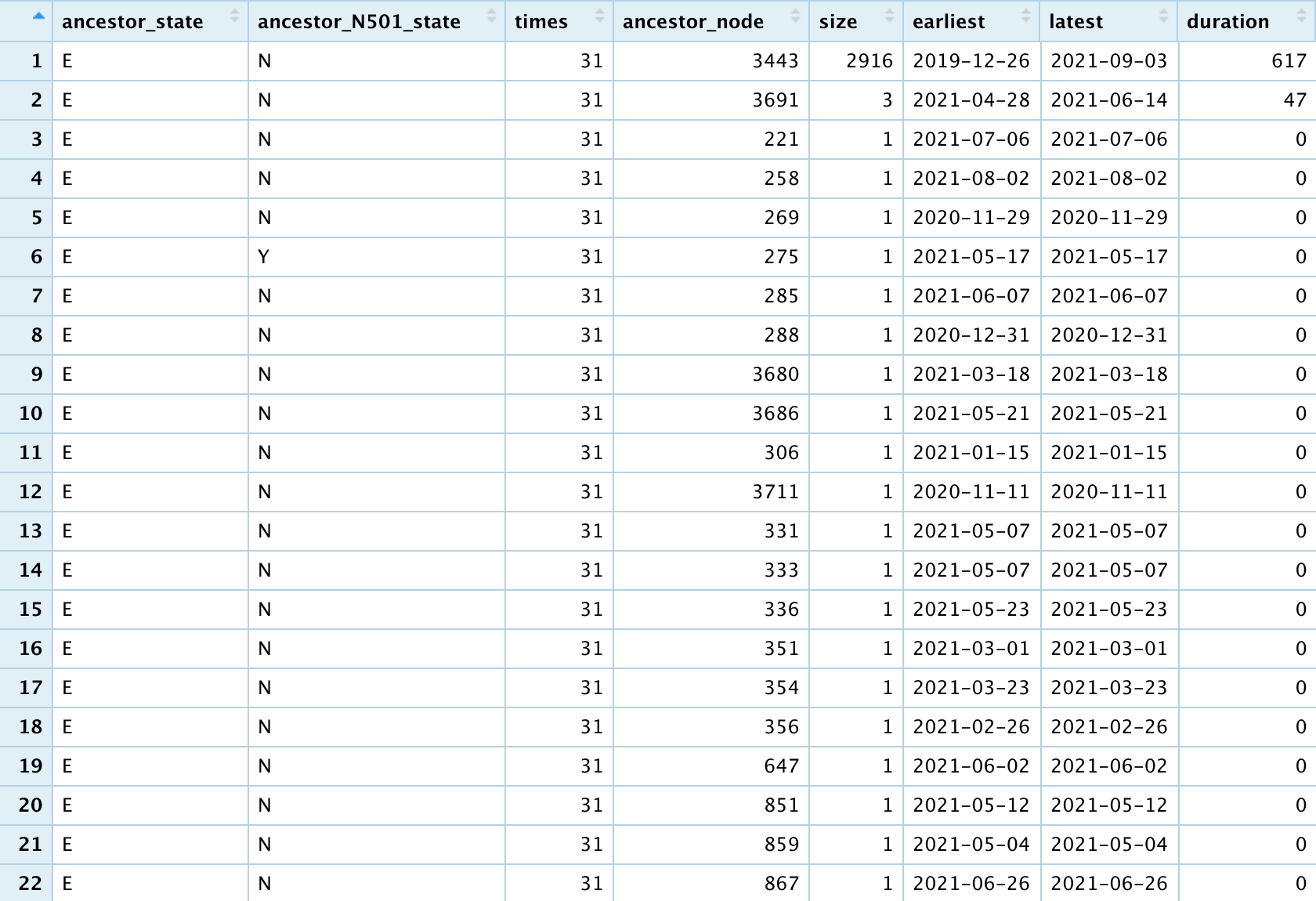
 

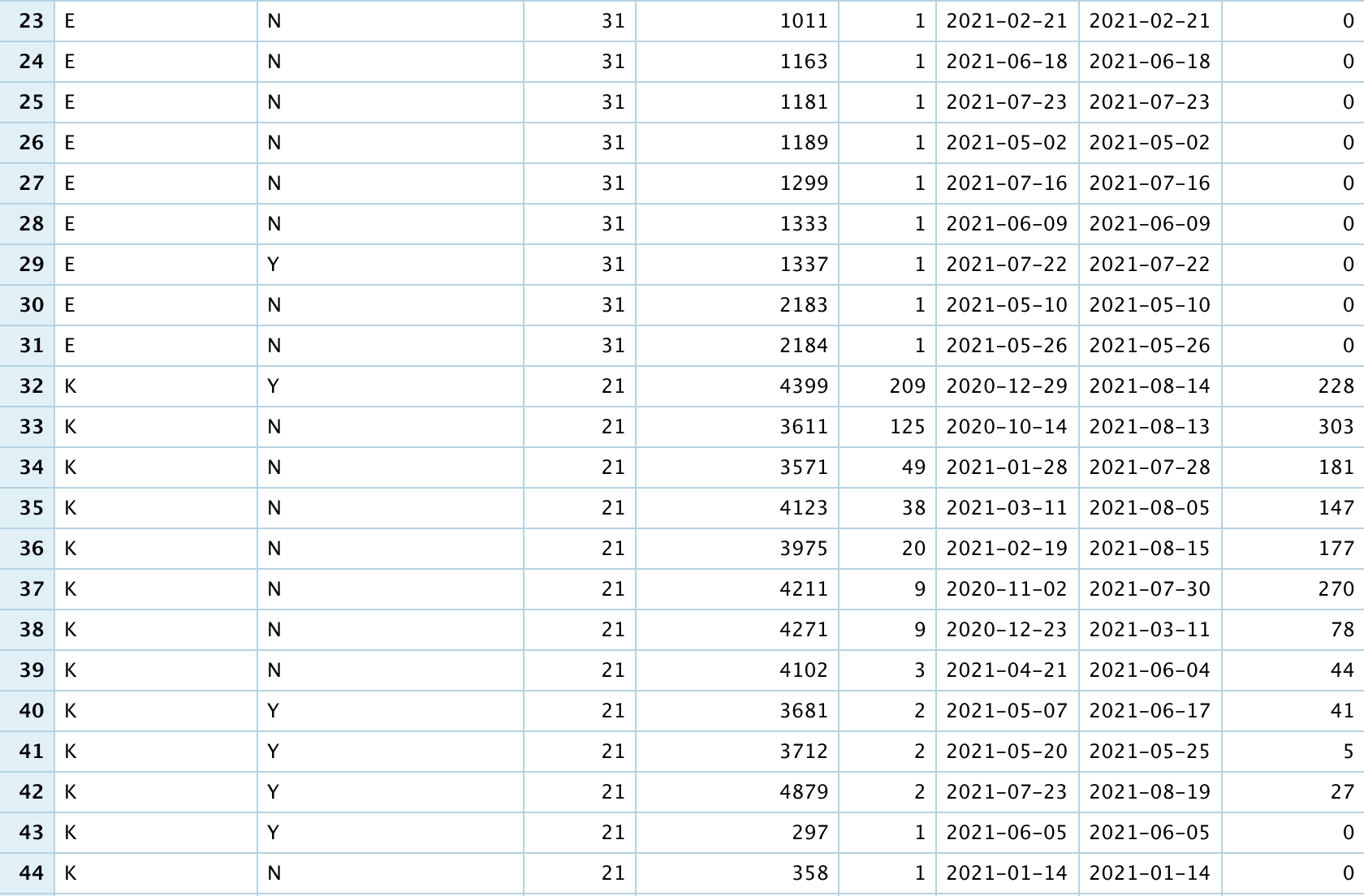
### E484-mutate clade emergence association with N501Y (epistasis)

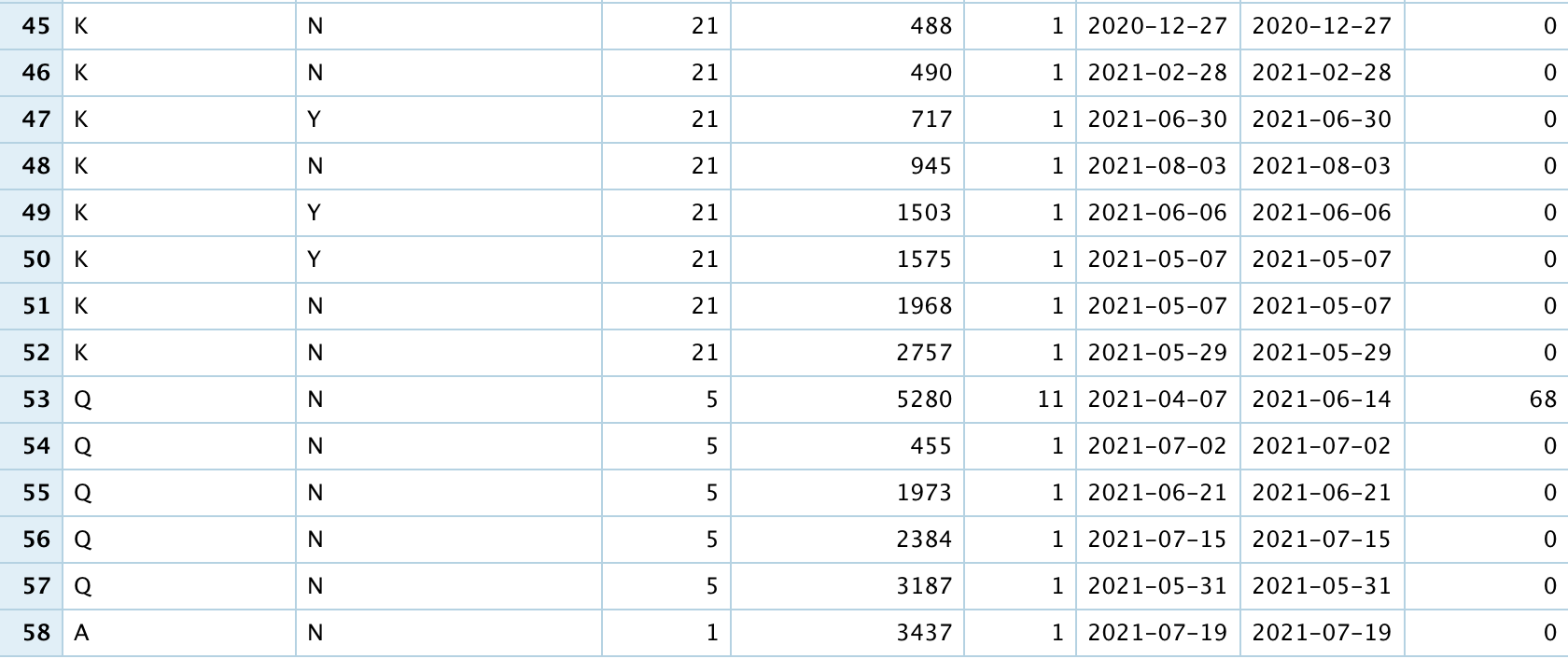
#### Proportion of N501 state at ancestral nodes of all E484 clades



data set “emerge”:



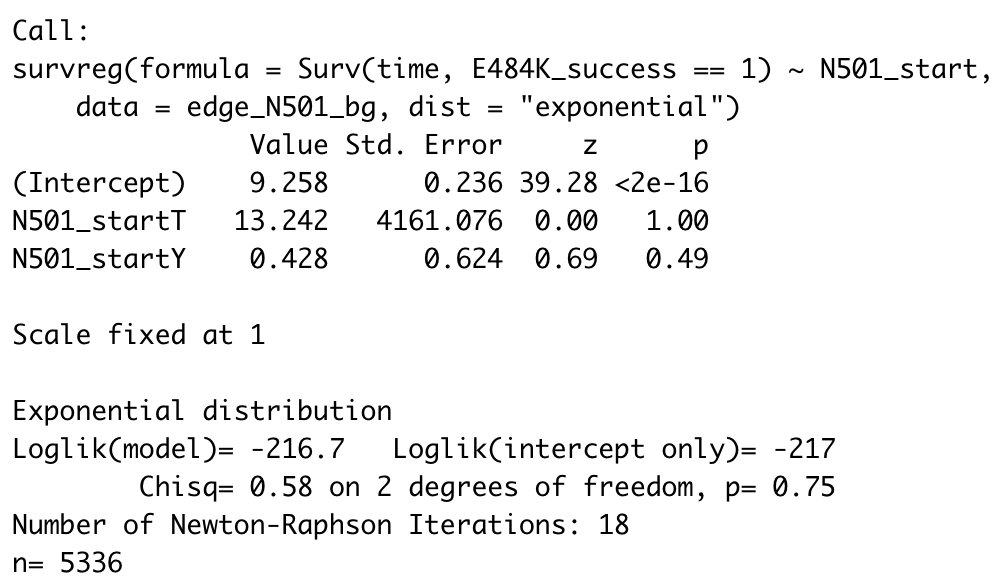


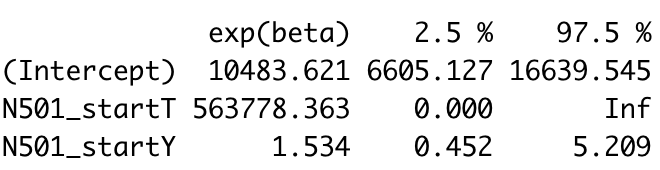


#### Logistic regression of E484K emergence by N501 state at ancestral node

#### Survival analysis of E484K emergence by N501 state at parent of ancestral node

**QUESTION: negative and zero branch lengths set to 0.0000000001???**

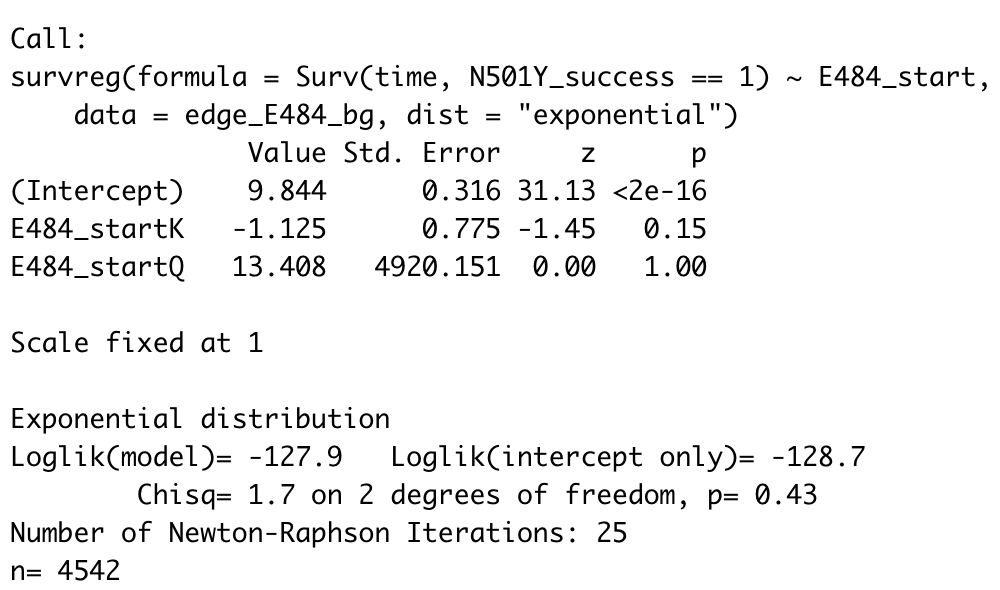


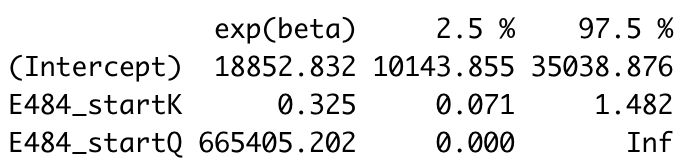


🡪 insig: if parent node has N501Y, longer survival time (E484K occurs slower)

🡪 if N501Y is already present, E484K is less favored?

#### Survival analysis of N501Y emergence by E484 state at parent of ancestral node





🡪 insig: if parent node has E484K, shorter survival time (N501Y occurs faster)

🡪 N501Y subsequently stabilizes E484K rather than promoting it to occur?