COVID-19 Spike Protein Phylogenetic trees

Ashvin Pidaparti, Bruno Bohn, Carrie Kistler, Kimberly Mandery

Background and Summary of Previous Findings

D614G mutation

Spike specific

Increased prevalence of mutations with higher replication rates and fitness

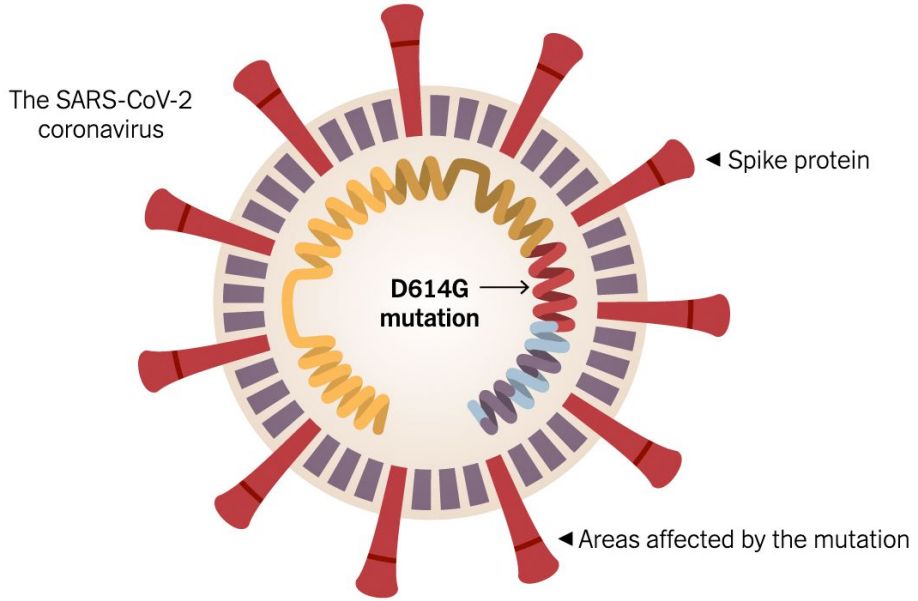


Image courtesy of the New York Times

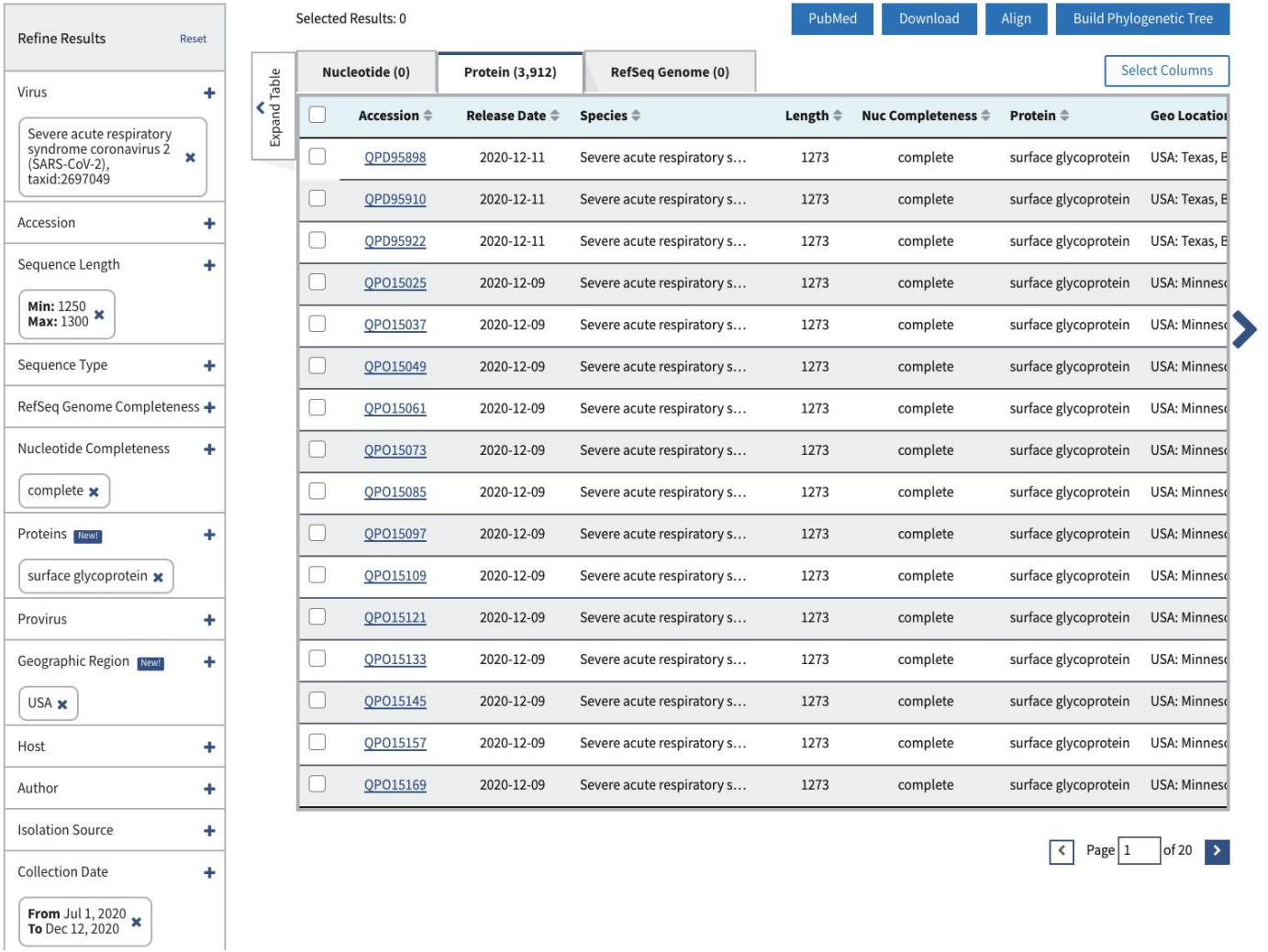
Sequences

Queried on **NCBI**

Spike protein (~1273 aa)

150 sequences:
- 15 states (10 each):
    - 5 before July
    - 5 July and after

Outgroup:
- 7 sequences from China, Jan/Feb 2020



Sequences

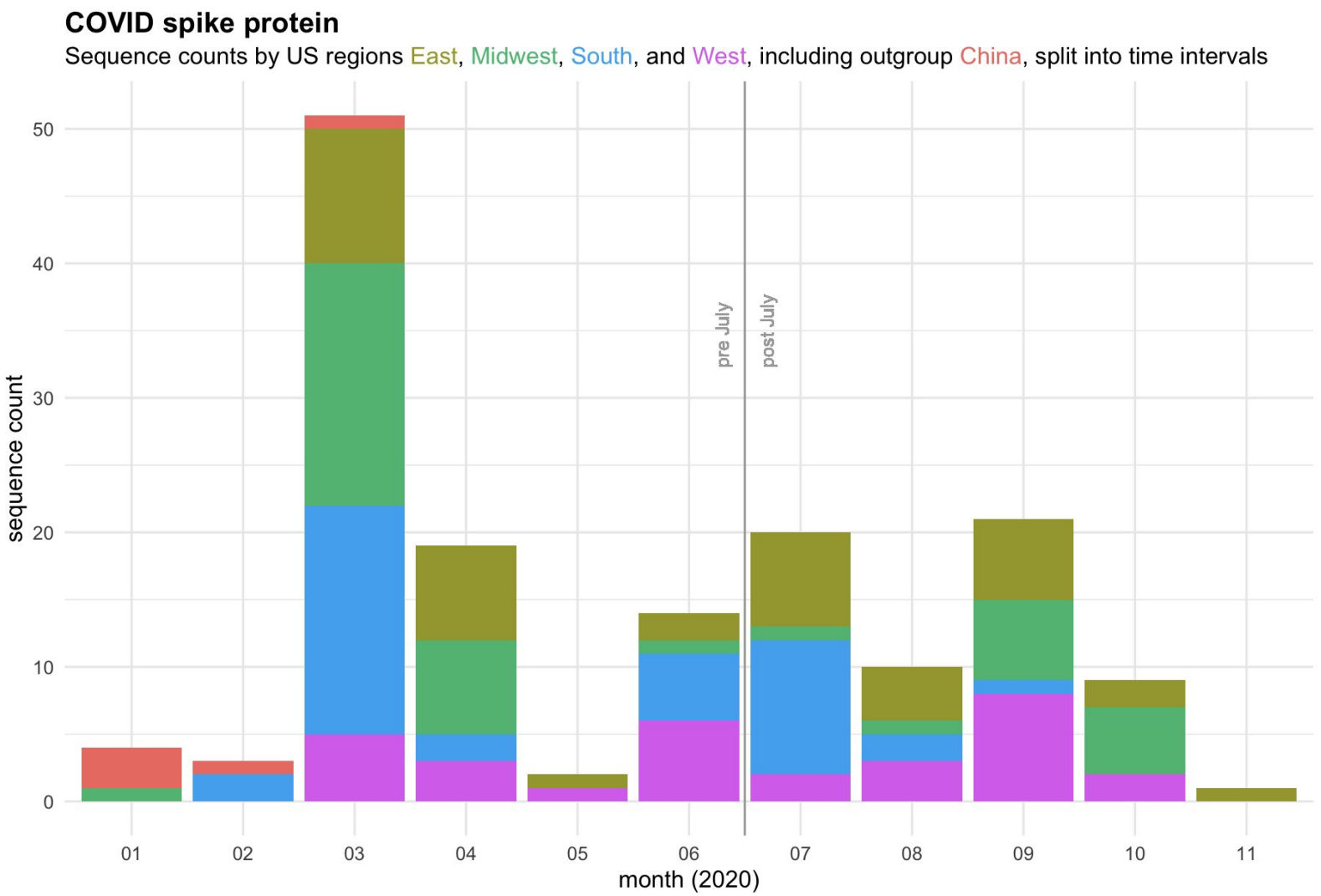
Aligned using **Muscle**

tried T-Coffee but was ineffective

Alignments | Result Summary | Phylogenetic Tree | Results Viewers | Submission Details

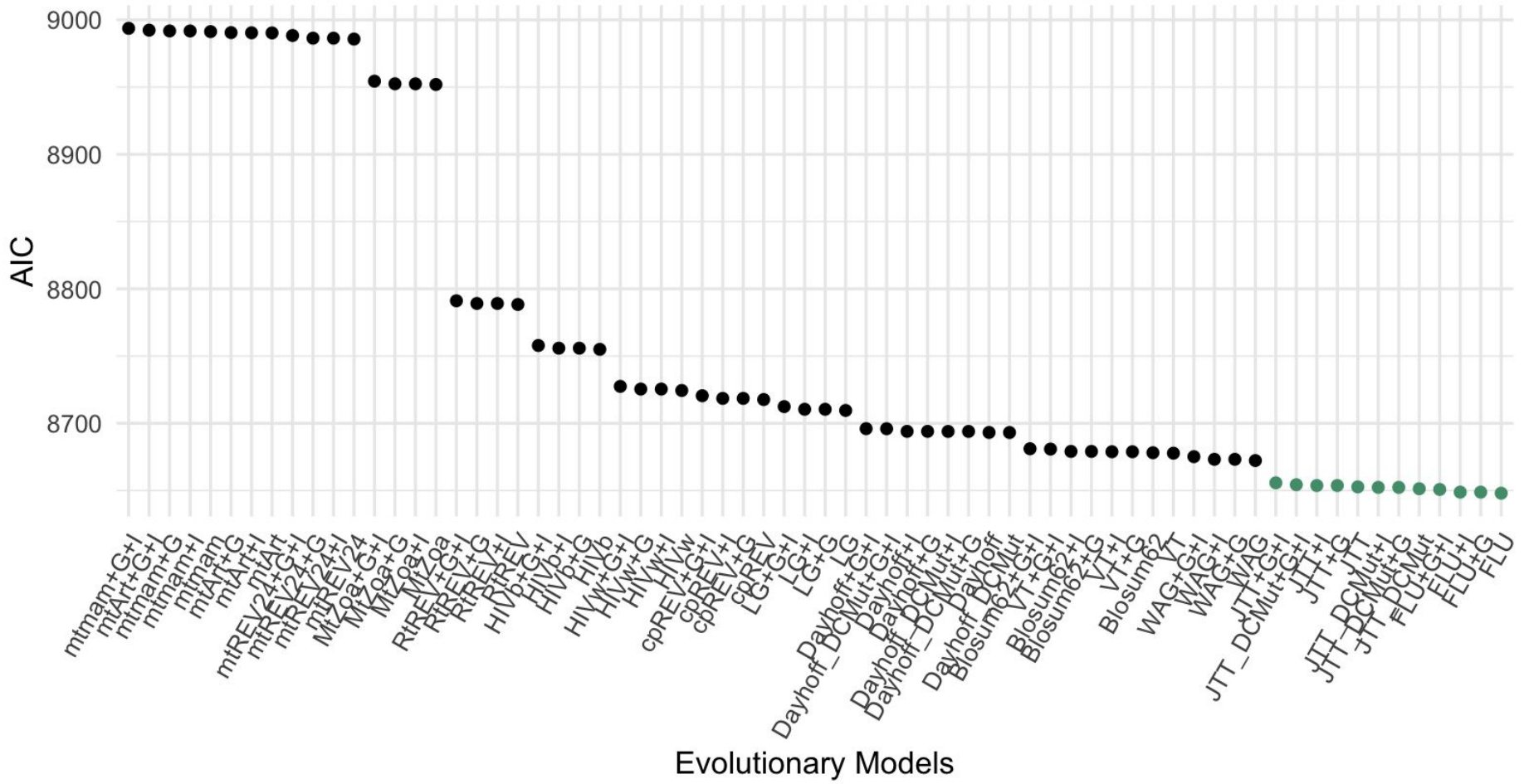
Download Alignment File | Show Colors

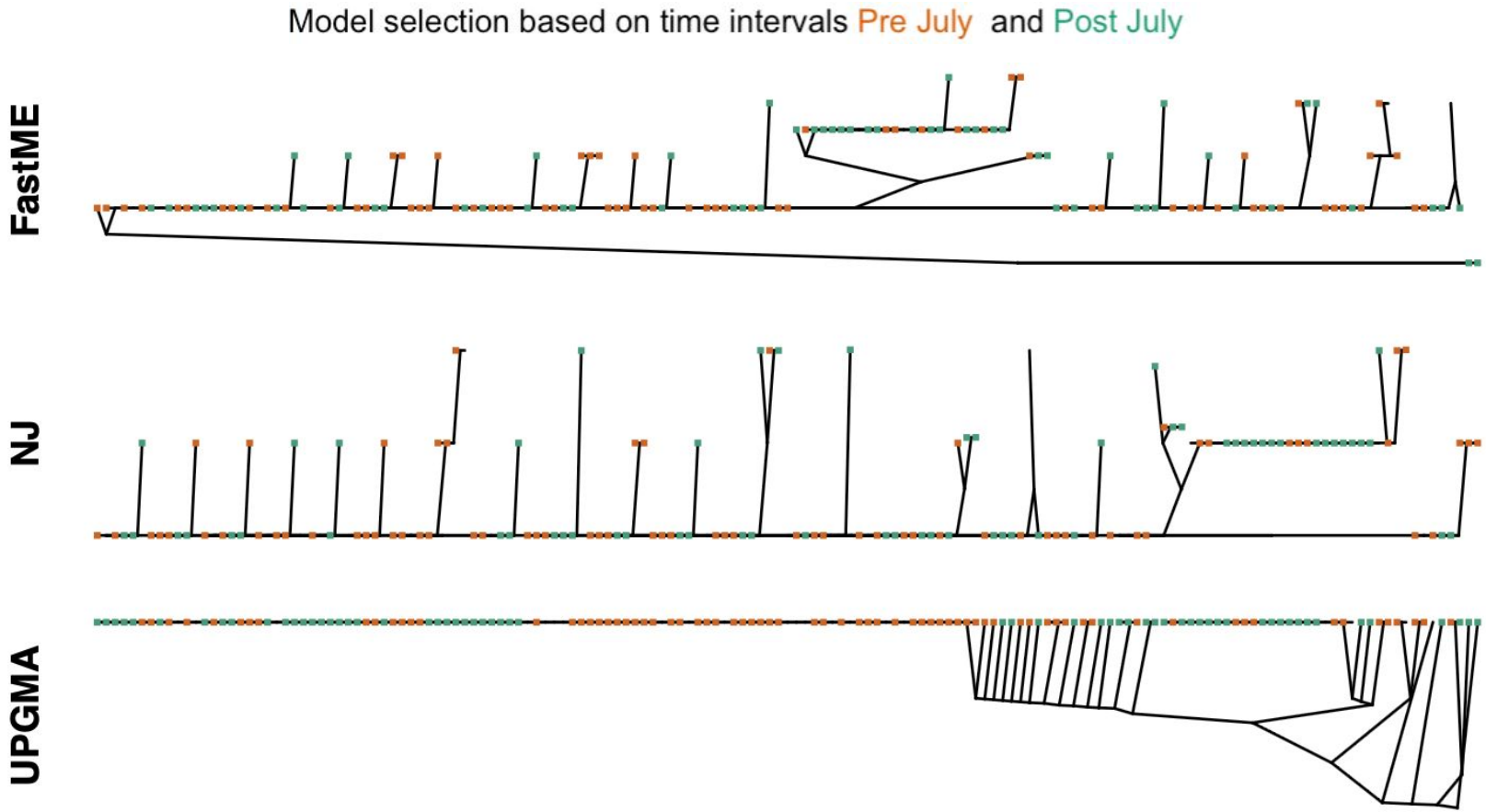
```
>QPL23245.1 |surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]|USA
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS
NVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIV
NNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLE
GKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQT
LLALHRSYLTPGDSSSGWXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRRRISN
CVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIAD
YNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC
NGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVN
FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP
GTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY
ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE
VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC
LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM
QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN
TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDP
LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL
QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD
SEPVLKGVKLHYT
>QPL23329.1 |surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]|USA
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS
NVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIV
```

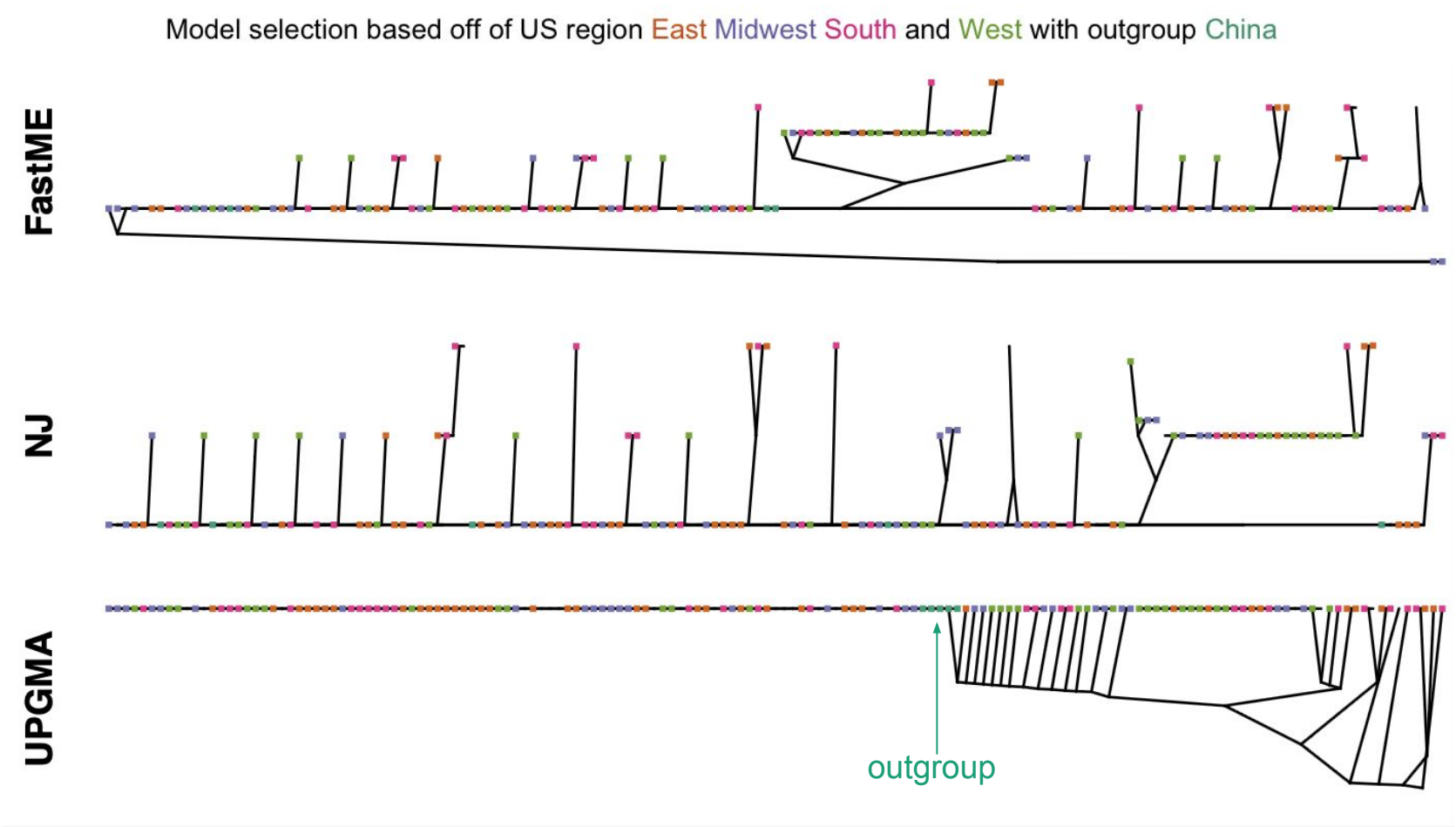

COVID spike protein
Sequence counts by US regions East, Midwest, South, and West, including outgroup China, split into time intervals
sequence count
0
10
20
30
40
50
pre July
post July
01
02
03
04
05
06
07
08
09
10
11
month (2020)

AIC measures

for finding best evolutionary models to be used in distance matrix calculation




Model selection based on time intervals Pre July and Post July
FastME
NJ
UPGMA


Model selection based off of US region East Midwest South and West with outgroup China
FastME
NJ
UPGMA
outgroup

Conclusions

No pattern in clustering was observed **by time period or US region.**

UPGMA suggested all **outgroup samples** (China, early 2020) to be **clustered with a large group of other samples.**

This result **suggests genetic variance largely occurred within the US**, without geographic limits.

A **larger sample size** and more **refined sample selection** methods **are warranted** to confirm findings.