COVID-19 Spike Protein Phylogenetic trees

Fall 2020

Background and Summary of Previous Findings

D614G mutation

Spike specific

Increased prevalence of mutations with higher replication rates and fitness

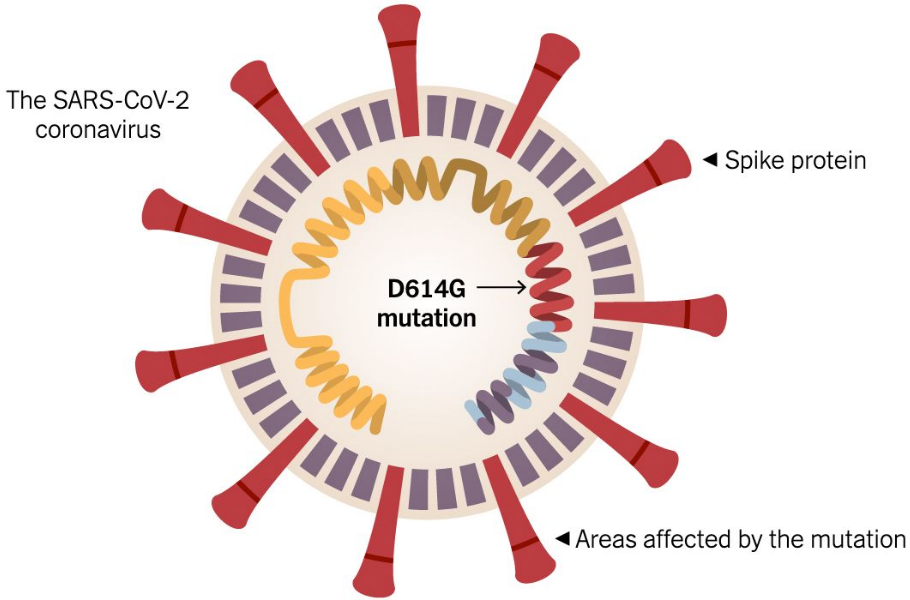


Image courtesy of the New York Times

Sequences

Queried on **NCBI**

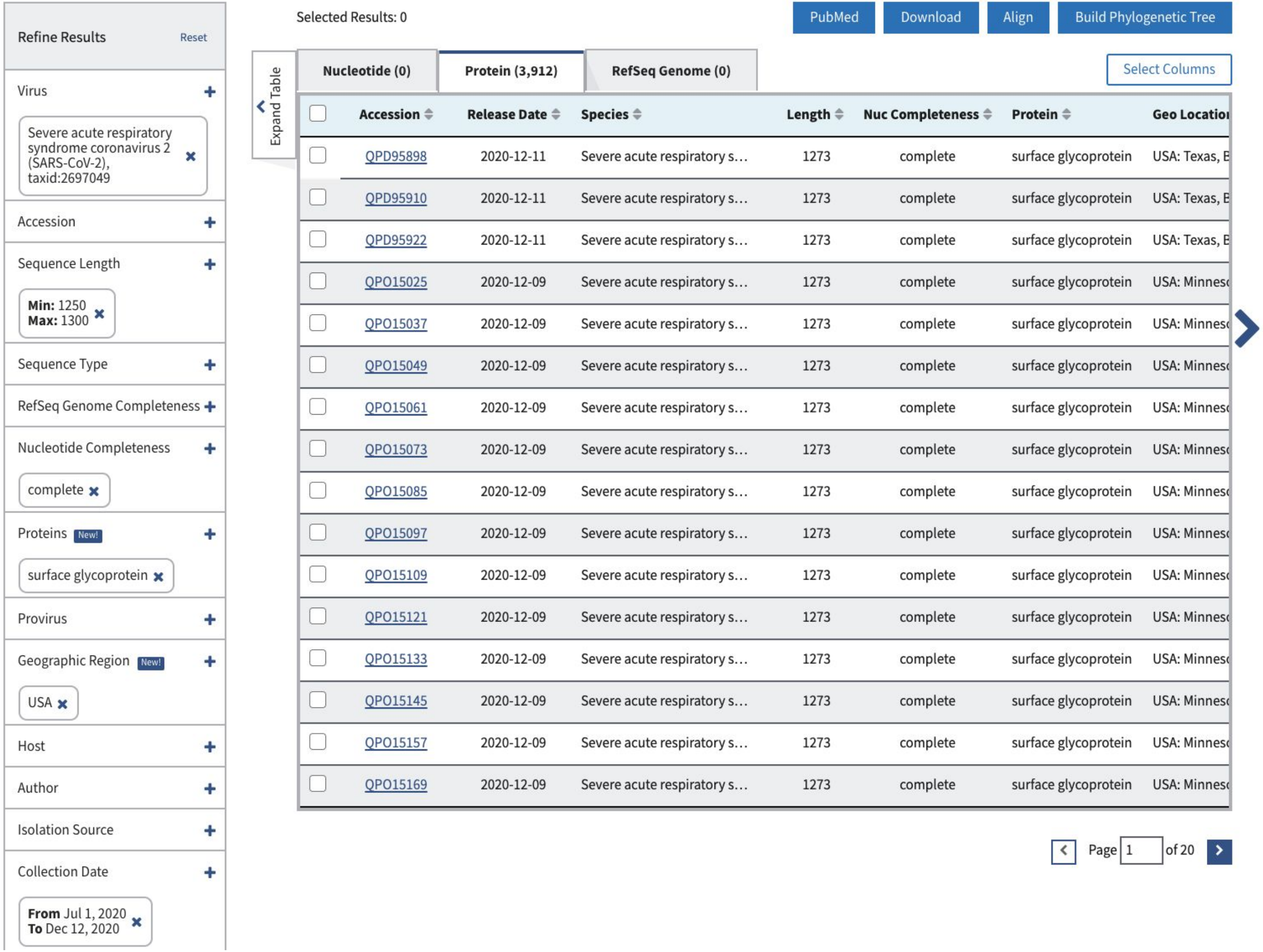
Spike protein (~1273 aa)

150 sequences:

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

Outgroup:

- 7 sequences from China, Jan/Feb 2020



Virus: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), taxid:2697049

Accession

Sequence Length: **Min:** 1250 **Max:** 1300

Sequence Type

RefSeq Genome Completeness

Nucleotide Completeness: complete

Proteins New!: surface glycoprotein

Provirus

Geographic Region New!: USA

Host

Author

Isolation Source

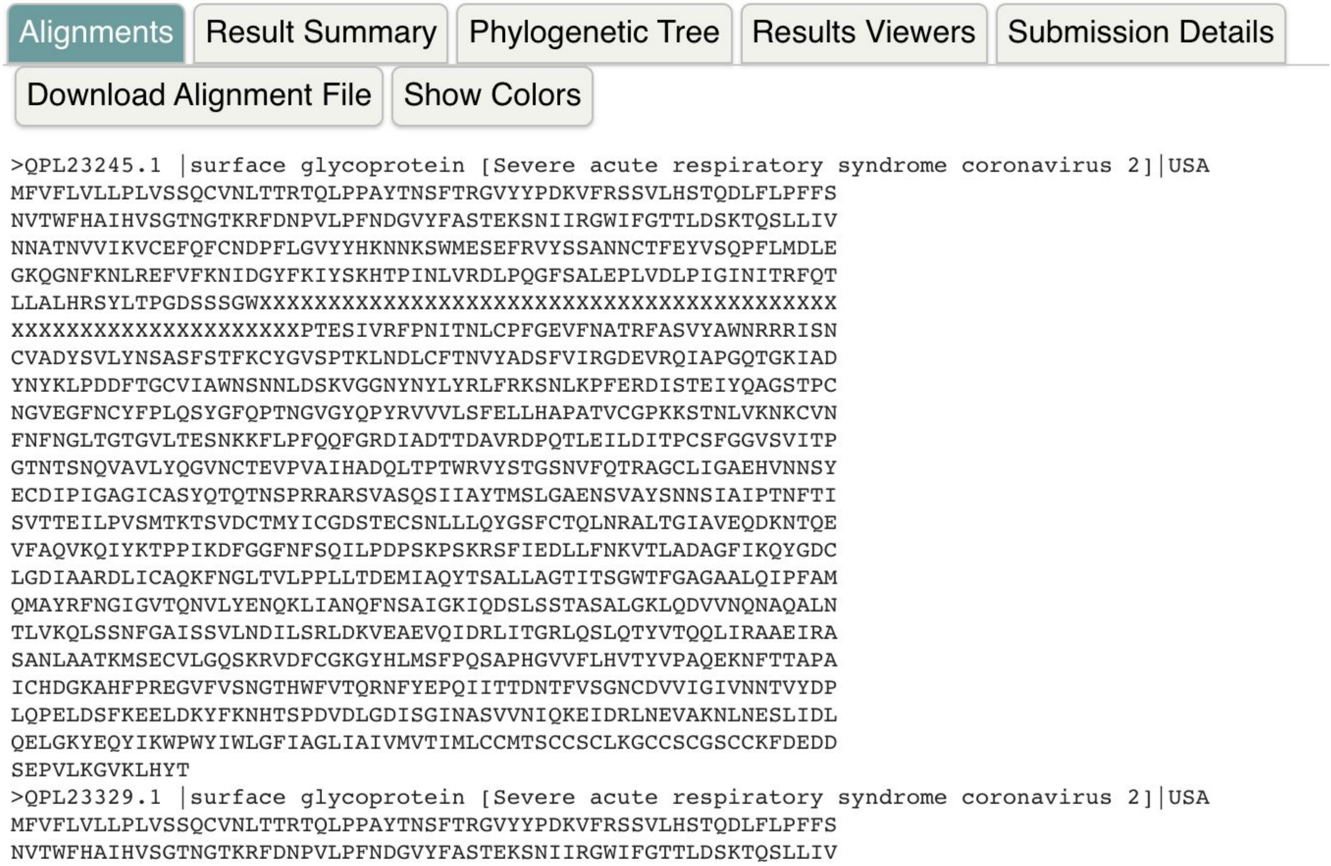
Collection Date: **From** Jul 1, 2020 **To** Dec 12, 2020

☐	Accession	Release Date	Species	Length	Nuc Completeness	Protein	Geo Locatio
☐	QPD95898	2020-12-11	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Texas, E
☐	QPD95910	2020-12-11	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Texas, E
☐	QPD95922	2020-12-11	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Texas, E
☐	QPO15025	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15037	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15049	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15061	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15073	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15085	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15097	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15109	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15121	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15133	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15145	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15157	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes
☐	QPO15169	2020-12-09	Severe acute respiratory s...	1273	complete	surface glycoprotein	USA: Minnes

Sequences

Aligned using **Muscle**

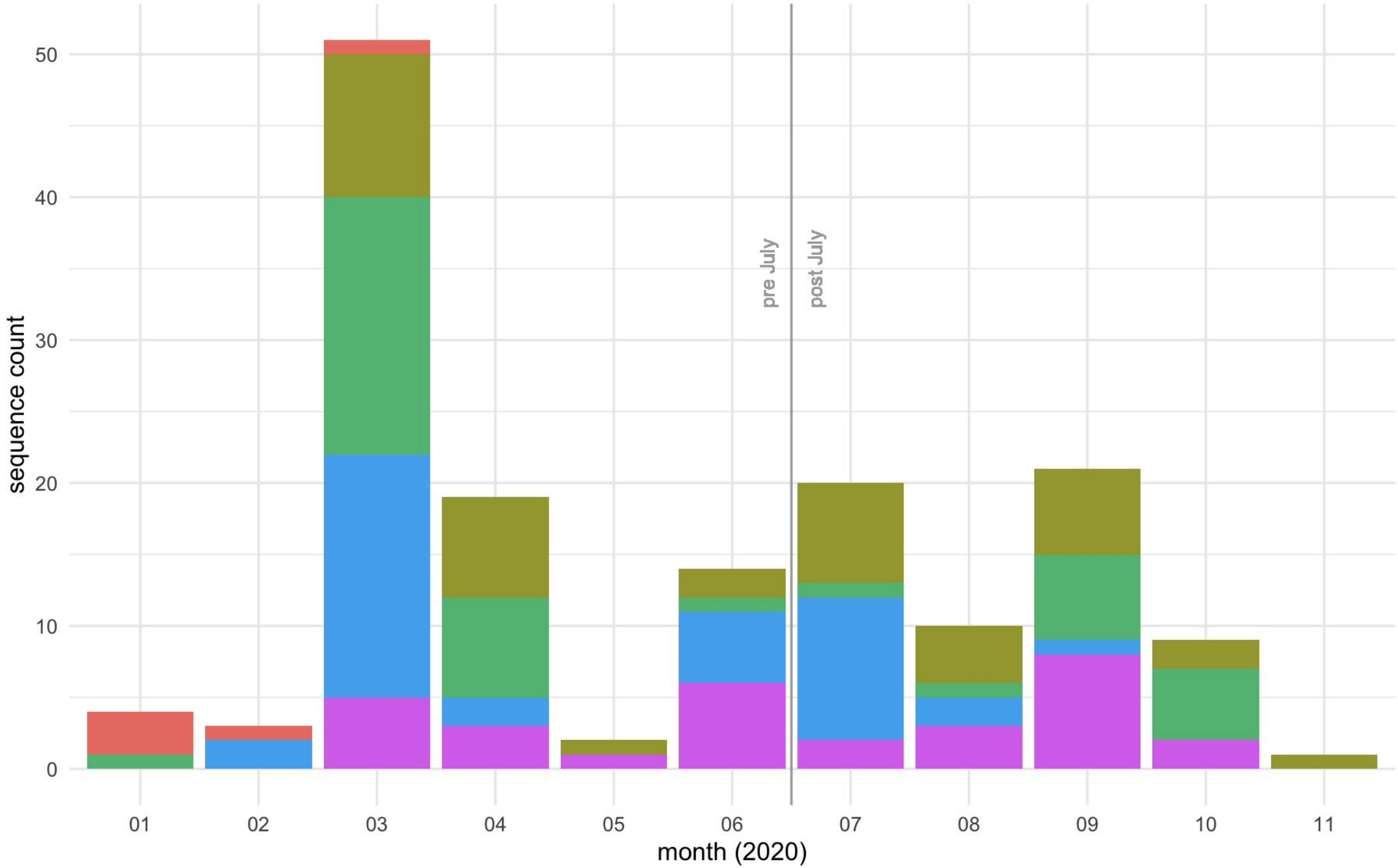
tried T-Coffee but was ineffective

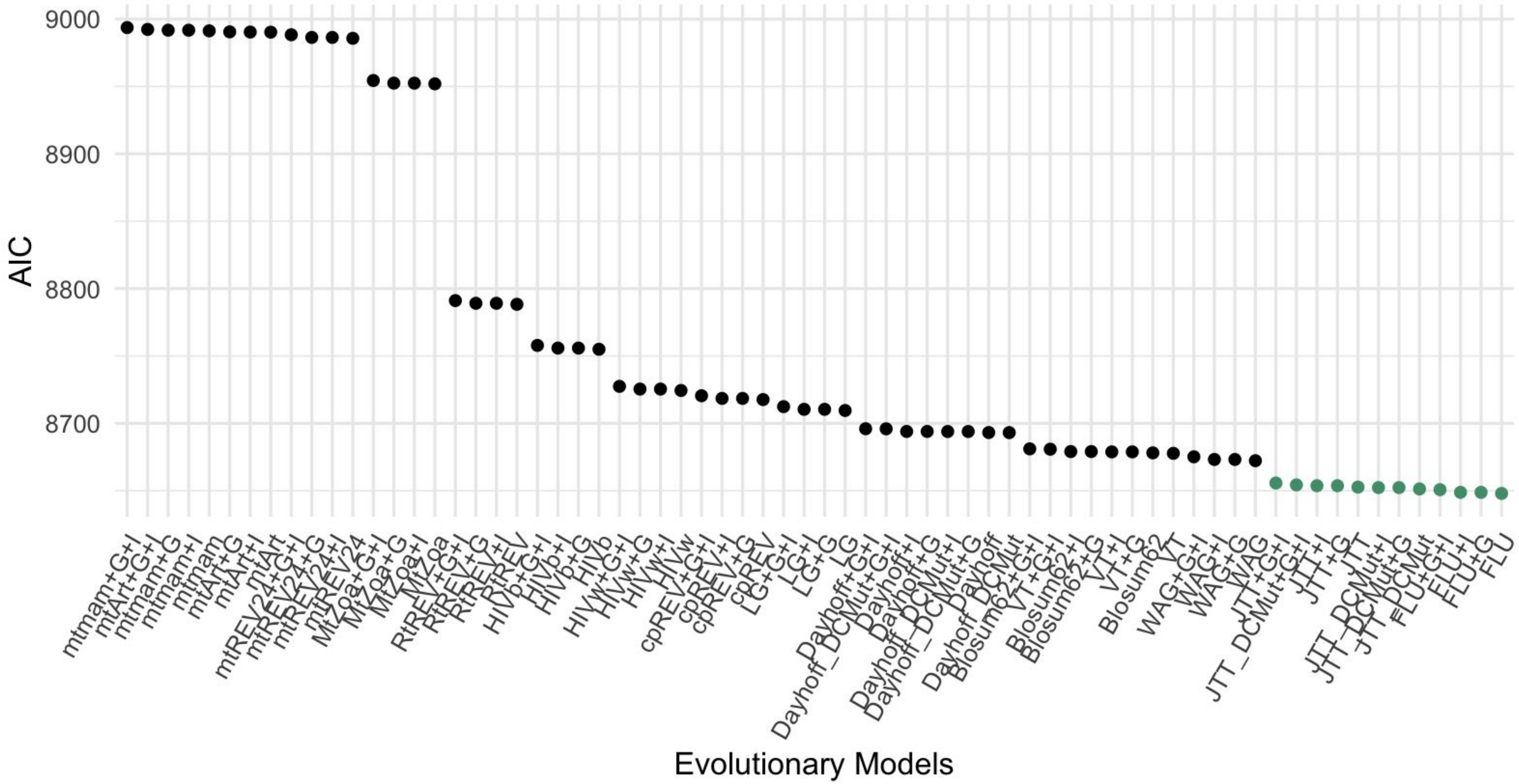


```
>QPL23245.1 |surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]|USA
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS
NVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIV
NNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLE
GKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQT
LLALHRSYLTPGDSSSGWXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRRRISN
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

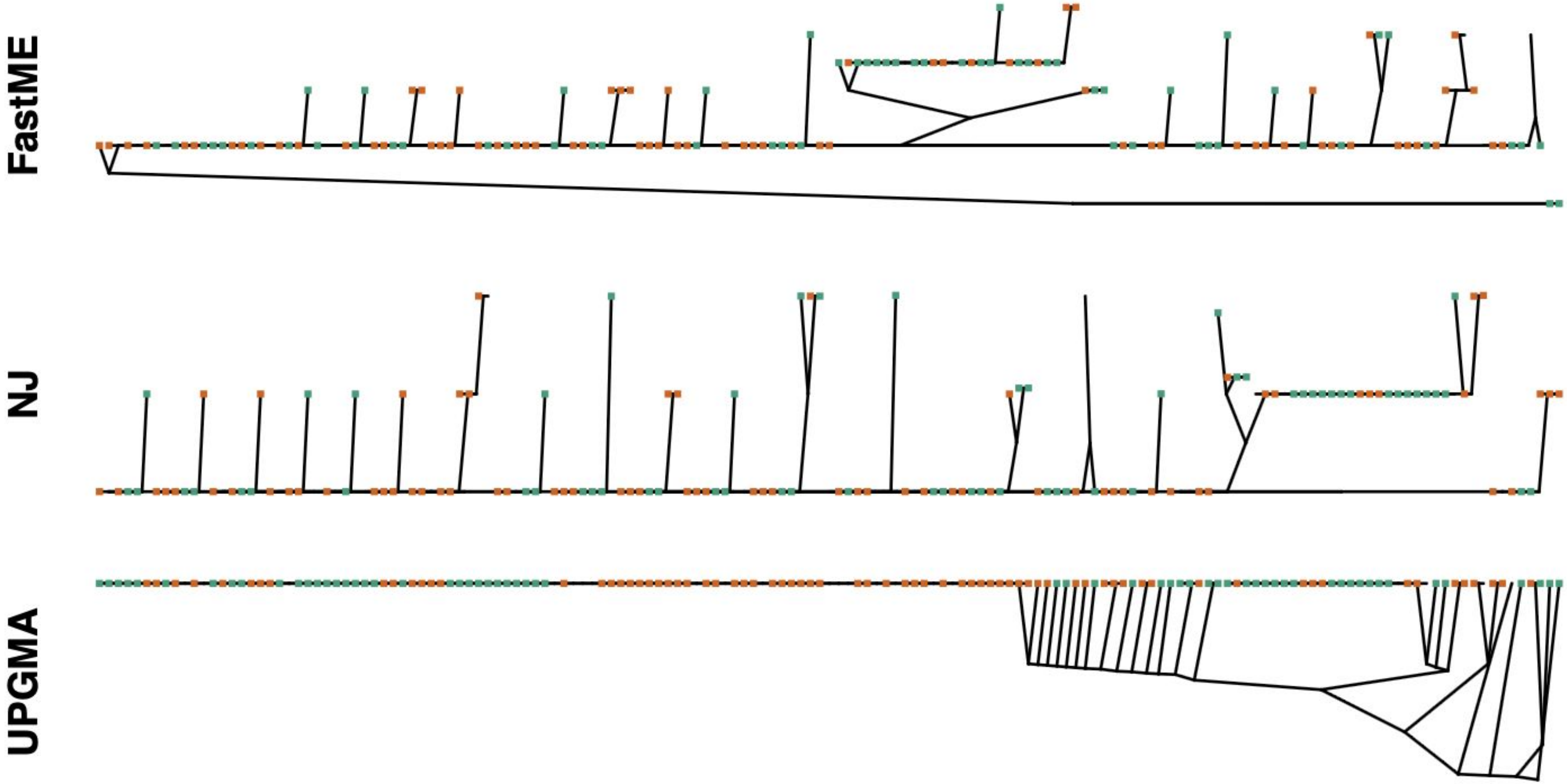




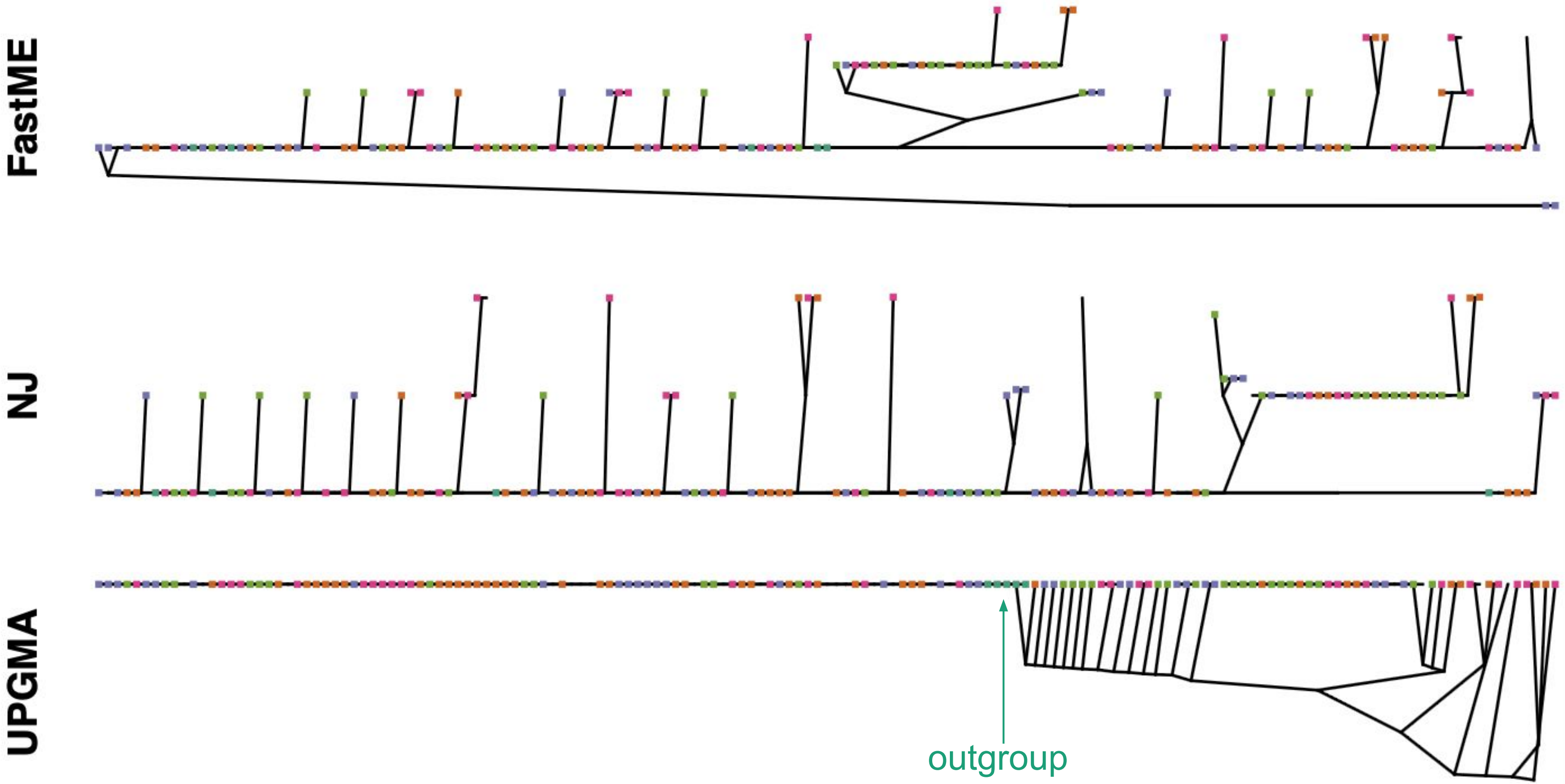
AIC measures

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

Model selection based on time intervals Pre July and Post July



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



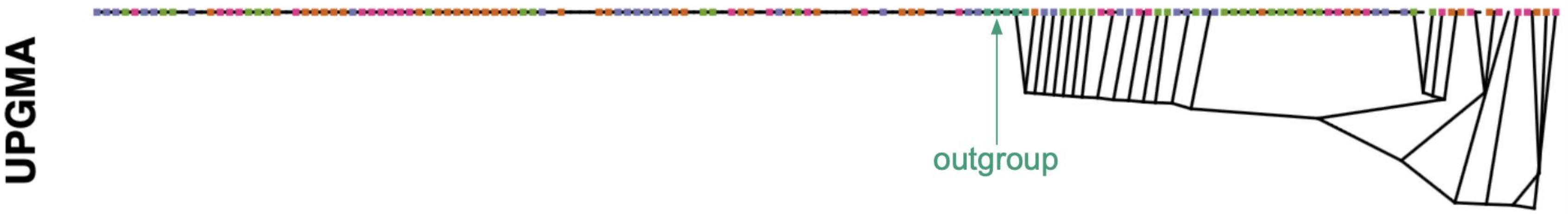
Conclusions

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.


UPGMA
outgroup