

December 2019

November 2020

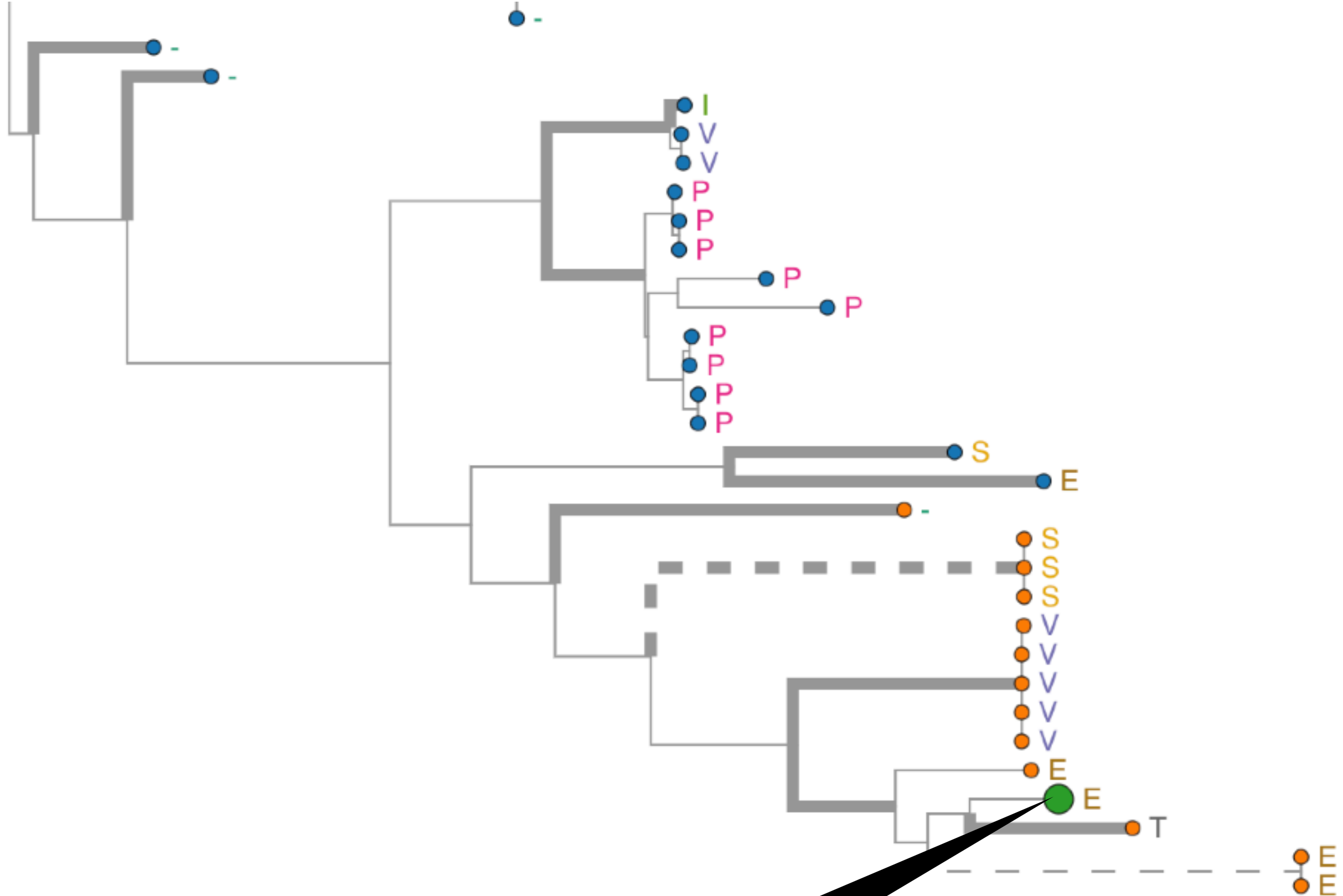
May 2021

- Use the evolutionary history in related *Sarbecoviruses* to predict which codons and amino-acids are “expected” in homologous SARS-CoV-2 positions.
- Our evolutionary model uses inferred site-level biochemical property importance to impute evolutionary credibility.



Continued evolution, complex
selection dynamics, transition to
endemic?

<https://observablehq.com/@spohn/visualizing-selection-analysis-results-for-evolution>



Predicting S/484 possible states
based on nCoV evolution

Evolutionary credibility report:

Codon	AA	Predicted probability in SARS-CoV-2
GAA	E	0.307
GAG	E	0.109
GTA	V	0.0818
GTT	V	0.0676
AAA	K	0.0477
GAT	D	0.0405
GCA	A	0.0315
GTG	V	0.0296
GGA	G	0.0272
GTC	V	0.0261
GAC	D	0.0195
AAG	K	0.0169
CAA	Q	0.0162
GCT	A	0.0154

Variable position => large admissible set of codons

December 2019

November 2020

May 2021

Continued evolution, complex selection dynamics, transition to endemic?

- Use the evolutionary history in related *Sarbecoviruses* to predict which codons and amino-acids are “expected” in homologous SARS-CoV-2 positions.

Evolutionary credibility report:

Codon	AA	Predicted probability in SARS-CoV-2
GAA	E	0.307
GAG	E	0.109
GTA	V	0.0818
GTT	V	0.0676
AAA	K	0.0477
GAT	D	0.0405
GCA	A	0.0315
GTG	V	0.0296
GGA	G	0.0272
GTC	V	0.0261
GAC	D	0.0195
AAG	K	0.0169
CAA	Q	0.0162
GCT	A	0.0154

- Our evolutionary model uses inferred site-level biochemical property importance and evolutionary credibility

Variable position => large admissible set of codons

Predicting S/484 possible states based on nCOV evolution

December 2019

November 2020

May 2021



For a given set of SARS-CoV-2 genomic sites compare predicted probabilities of finding specific codons at given genomic sites (based on the [evolutionary analysis of closely related animal sarbecoviruses](#)) vs observed variation with a median of **3078961.5** consensus genomes per codon of SARS-CoV-2 from GISAID.

[Download .JSON data](#)

9440 genomic codon loci analyzed	...	8688 loci with observed variants	↔	8723 loci with evolutionary predictions	📖
4:8 Median (95%) observed codon variants per locus	🔗	4:8 Median (95%) predicted codon variants per locus	🎯	0 Median (predicted-observed) count difference per locus	📄
8753 Loci with all variants at ≥ 0.1% predicted	👍	1491 Loci with unpredicted variants at ≥ 0.01%	👎	724 Sites with perfectly predicted minority variants at 0.01% threshold	📊
0.356 Spearmankman rank correlation between observed and predicted site entropies	📈	0.274 Fraction of variable loci where the top minority codon was correctly predicted	📈	1.74e+4 Prediction bit-score compared to a matched complexity random model (0.01% threshold)	🕒