What can sequence-based analyses of natural selection of SARS-CoV-2 tell us about the past, present, and future evolution of the virus?

Insights from selection analysis of complete genomes

COVID-19 Virtual Symposium - October 20, 2021

Sergei L. Kosakovsky Pond
Professor of Biology
Institute for Genomics and Evolutionary Medicine @ Temple University

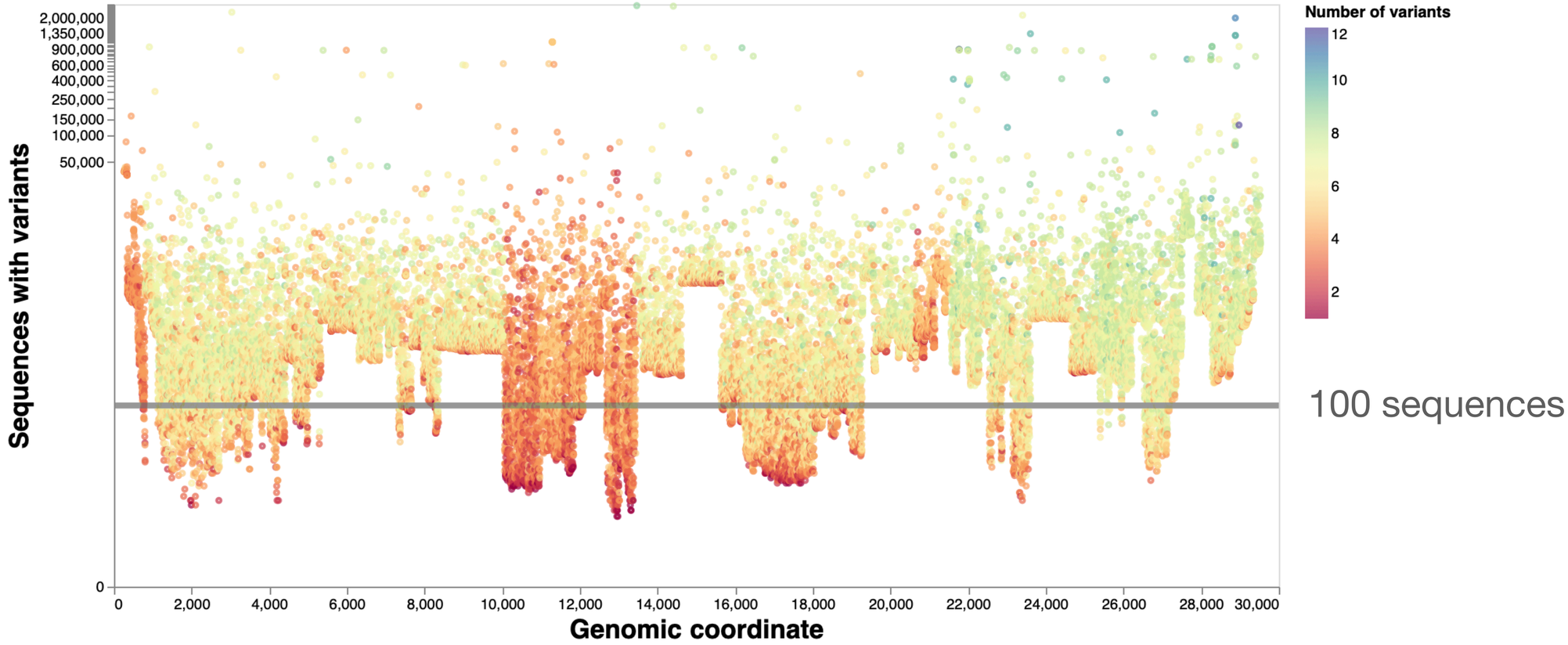
spond@temple.edu
http://lab.hyphy.org
github.com/spond
@sergeilkp

Key drivers of adaptation in rapidly evolving pathogens

- Zoonoses and transmission to new hosts (both species and individuals)
- Immune selection (CTL, innate, antibody)
- Development of drug resistance
- Virulence/transmissibility
- Host/pathogen arms-races, e.g. host antiviral factors
- **Most** of the time, **most** of the viral genome is **conserved**
- **Most** of the **observed variation** is **“neutral”**
- Changes that are **not neutral** are important to **detect early**, and, ideally, **predict**

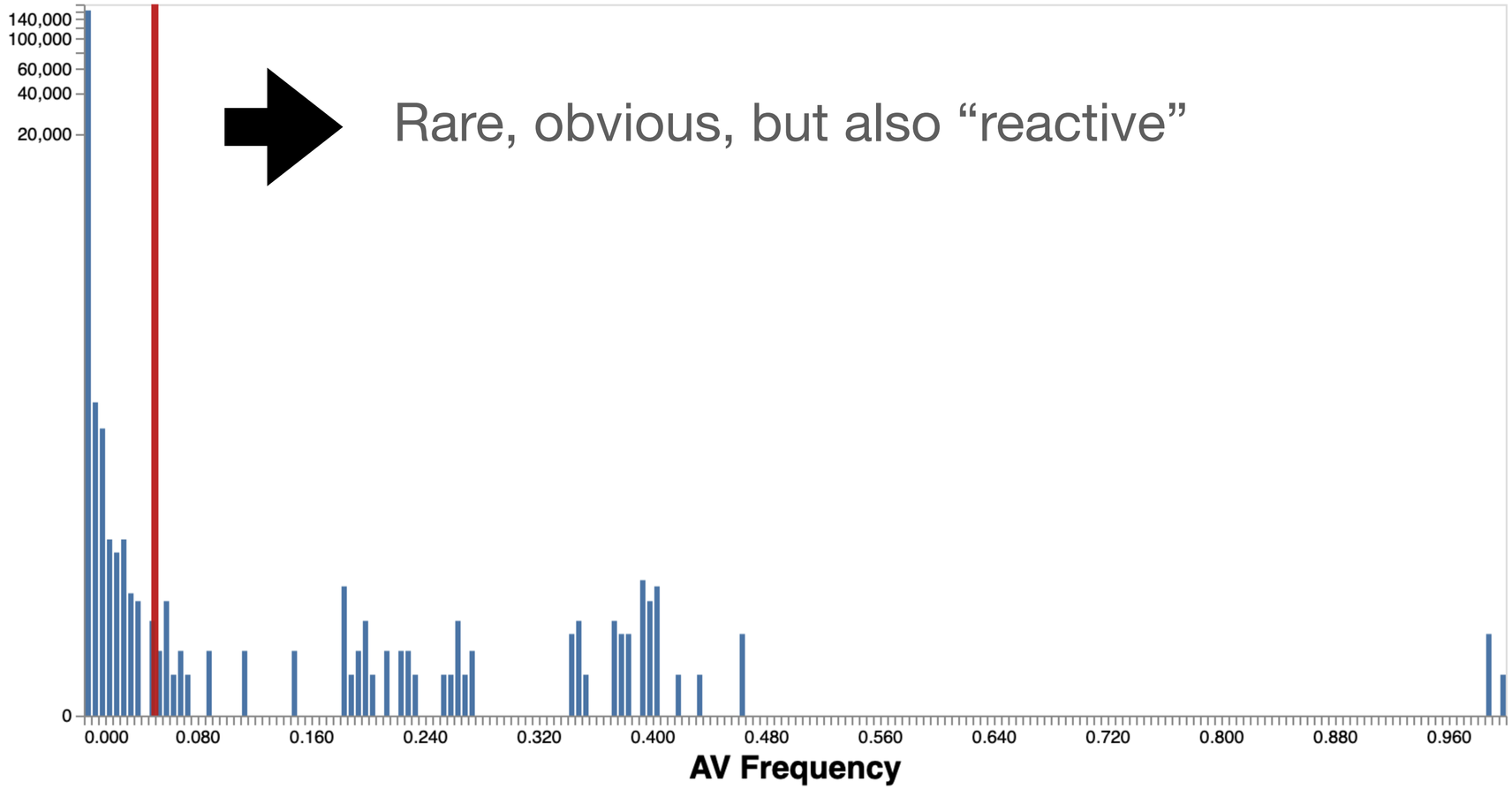
How much genomic variation is there in SARS-CoV-2?

- There are >4M genomes in GISAID and >1.3M NGS datasets in SRA for SARS-CoV-2
- At this point, nearly every genomic position has sequences with multiple allelic variants (AVs)



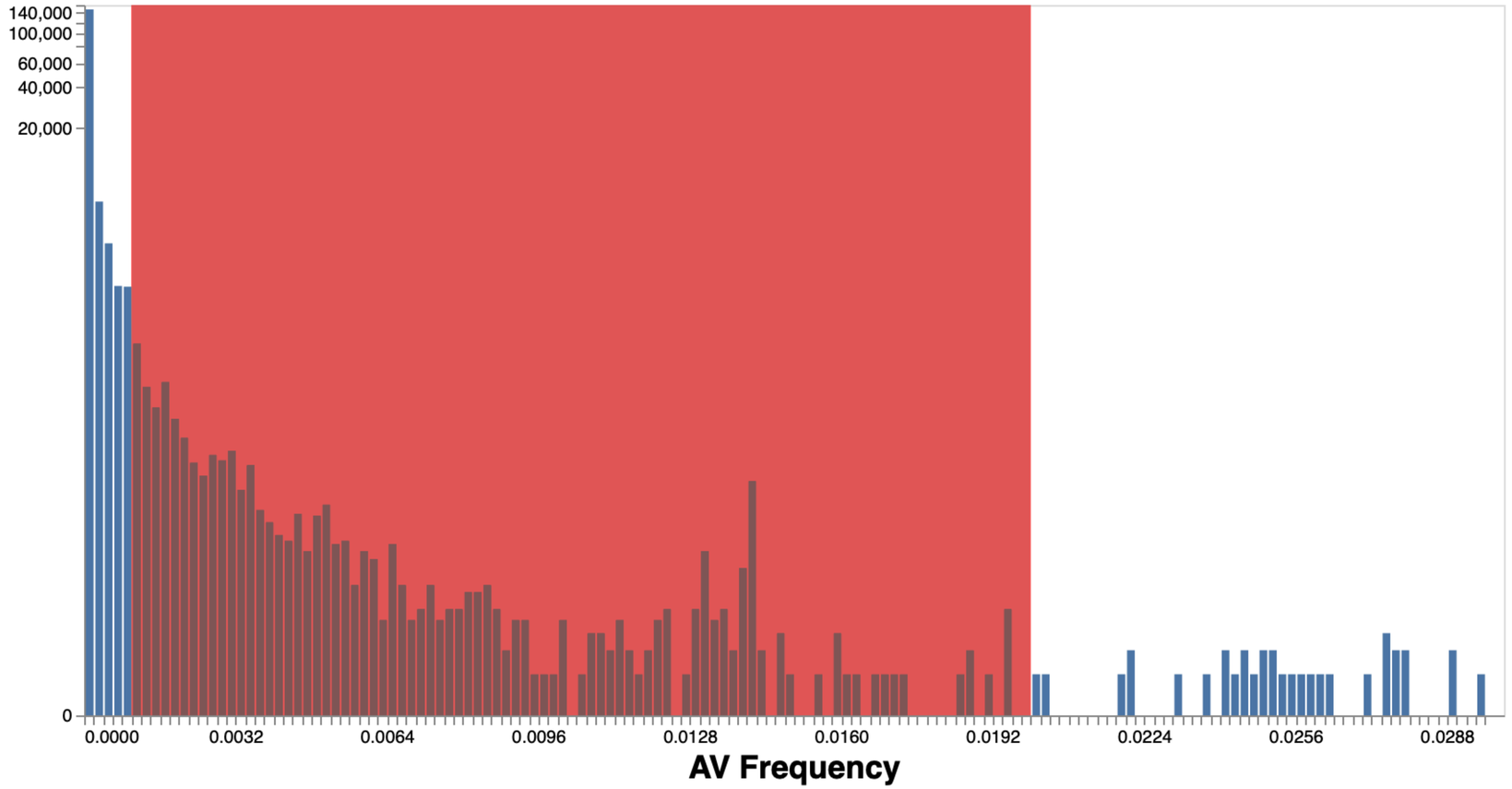
Which AVs are interesting and non-obvious?

- Most variation exists at low (<1%) frequency.
- Variants more frequent than 5% are rare (~100) and can be exhaustively tested.



Which AVs are interesting and non-obvious?

- Less common variants (e.g. 0.1% - 2%) are potentially interesting and non-obvious
- Too many (>2000) to test exhaustively



Analytical framework

- Key mutations have significant phenotypic and epidemiological significance.
- Analysis of substitution patterns in large sets of SARS-CoV-2 data can reveal genomic sites subject to **selective pressure => possibly important**
- Used a suite of stock and modified **dN/dS methods** for **coding sequence evolution** developed over ~2 decades and extensively used in other RNA viruses
- The methods are implemented in the HyPhy software package (www.hyphy.org)
- Our public servers (datamonkey.org) have processed >10,000 complex CoV evolutionary analyses from researchers worldwide.
- Spent considerable effort to scale up analyses and develop data reduction techniques to manage data sizes.
- Collaborative efforts, many other open source tools at covid-19.galaxyproject.org

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.

- Did SARS-CoV-2 experience selective pressure during its emergence from an animal progenitor?

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.

- Did SARS-CoV-2 experience selective pressure during its emergence from an animal progenitor?

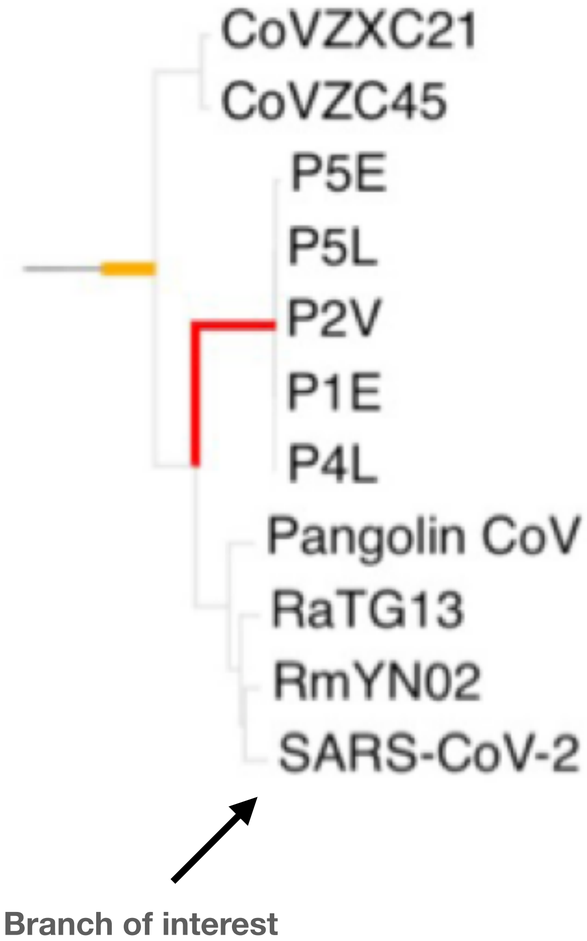
- Collected all closely related sarbecovirus sequences, and defined the nCOV clade of isolates including the reference SARS-CoV-2 strain.

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.

December 2019 | November 2020 | May 2021

Zoonosis/early adaptation

- Did SARS-CoV-2 experience selective pressure during its emergence from an animal progenitor?
- Collected all closely related sarbecovirus sequences, and defined the nCOV clade of isolates including the reference SARS-CoV-2 strain.
- Focused on selection and recombination history of the nCOV clade.



MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.

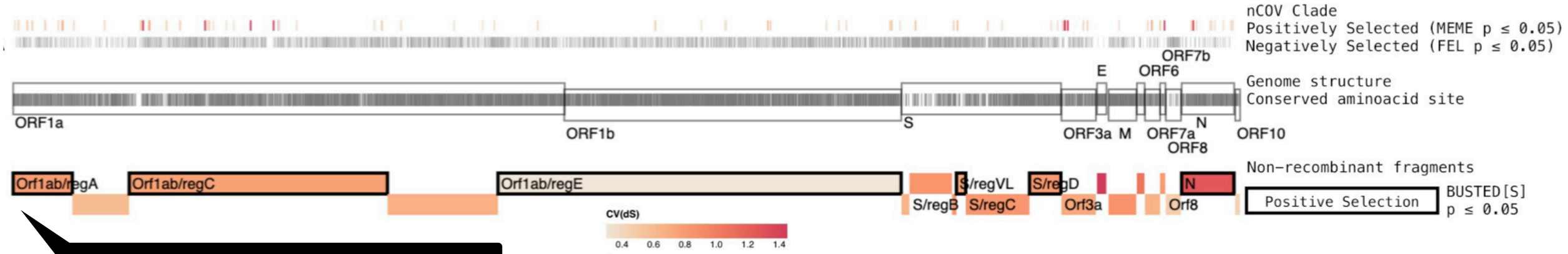

December 2019
November 2020
May 2021
Zoonosis/early adaptation

- Found evidence of extensive recombination history in the nCOV clade.
  - Later studies show evidence of cold/hot spots as well *Lytras et al 2021 BioRxiv*

December 2019 | November 2020 | May 2021

Zoonosis/early adaptation

- Found evidence of extensive recombination history in the nCOV clade.
  - Later studies show evidence of cold/hot spots as well *Lytras et al 2021 BioRxiv*

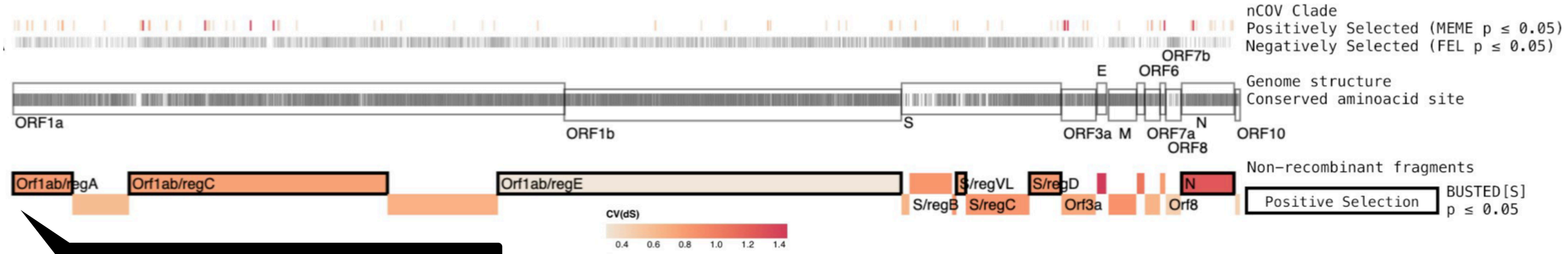


Many recombination breakpoints

December 2019 | November 2020 | May 2021

Zoonosis/early adaptation

- Found evidence of extensive recombination history in the nCOV clade.
  - Later studies show evidence of cold/hot spots as well *Lytras et al 2021 BioRxiv*
- Compelling argument for CG depletion as an adaptive trait.

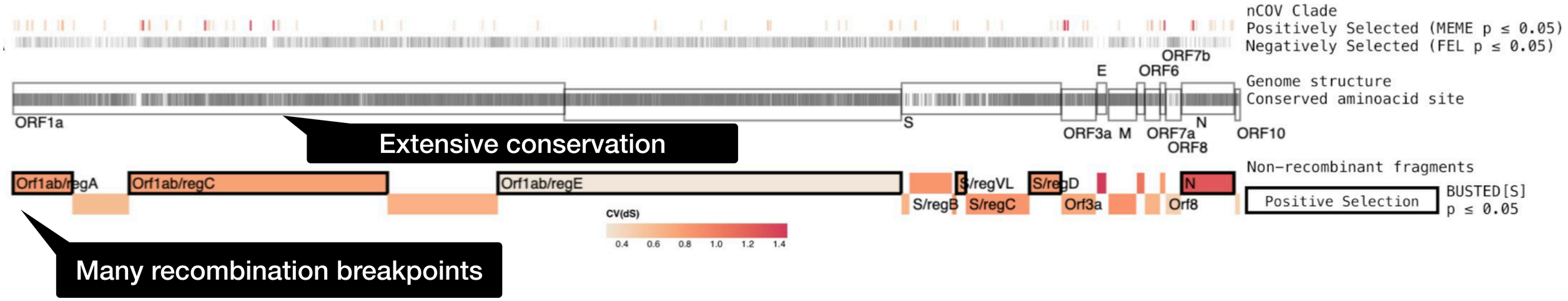


Many recombination breakpoints

December 2019 | November 2020 | May 2021

Zoonosis/early adaptation

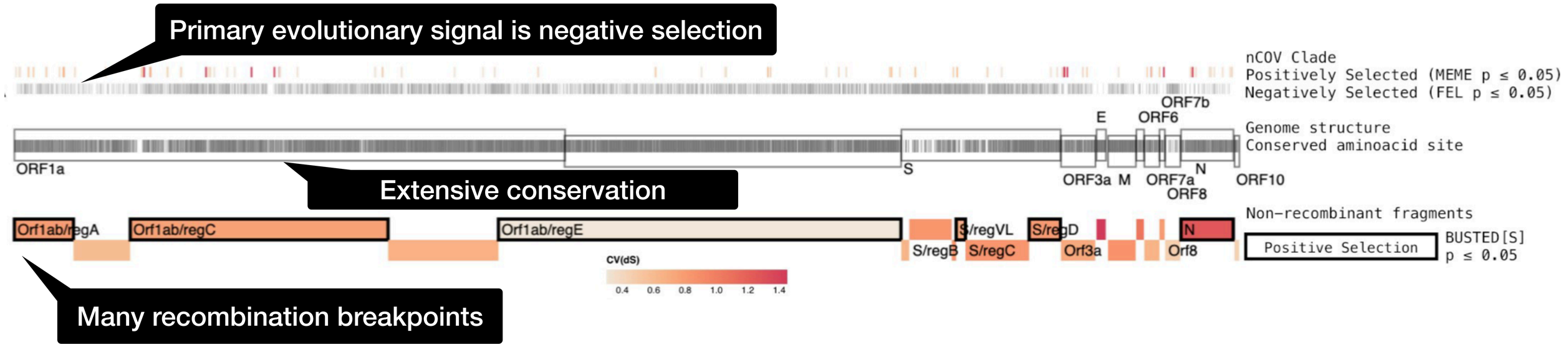
- Found evidence of extensive recombination history in the nCOV clade.
  - Later studies show evidence of cold/hot spots as well *Lytras et al 2021 BioRxiv*
- Compelling argument for CG depletion as an adaptive trait.



December 2019 | November 2020 | May 2021

Zoonosis/early adaptation

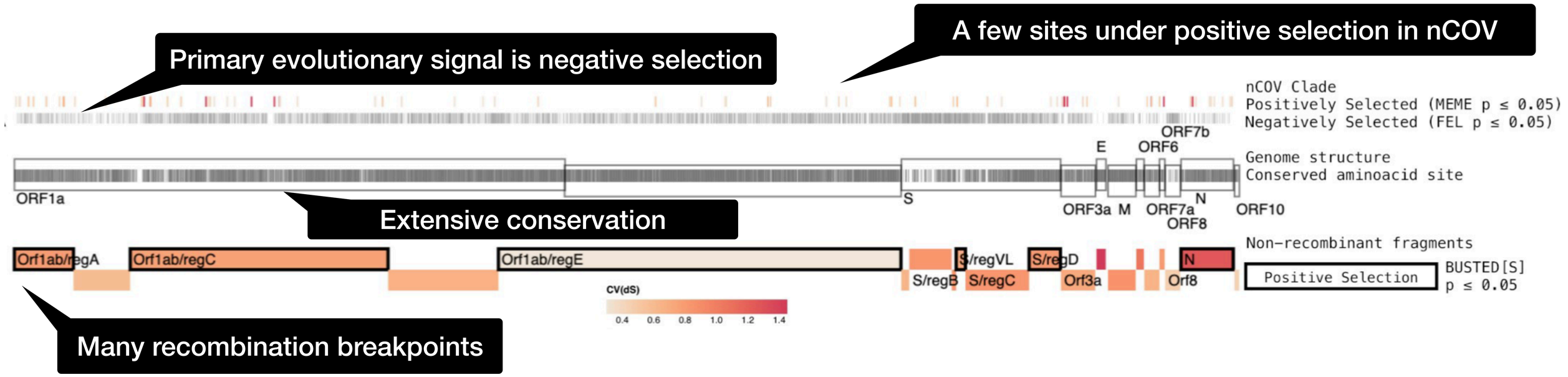
- Found evidence of extensive recombination history in the nCOV clade.
  - Later studies show evidence of cold/hot spots as well *Lytras et al 2021 BioRxiv*
- Compelling argument for CG depletion as an adaptive trait.



December 2019 | November 2020 | May 2021

Zoonosis/early adaptation

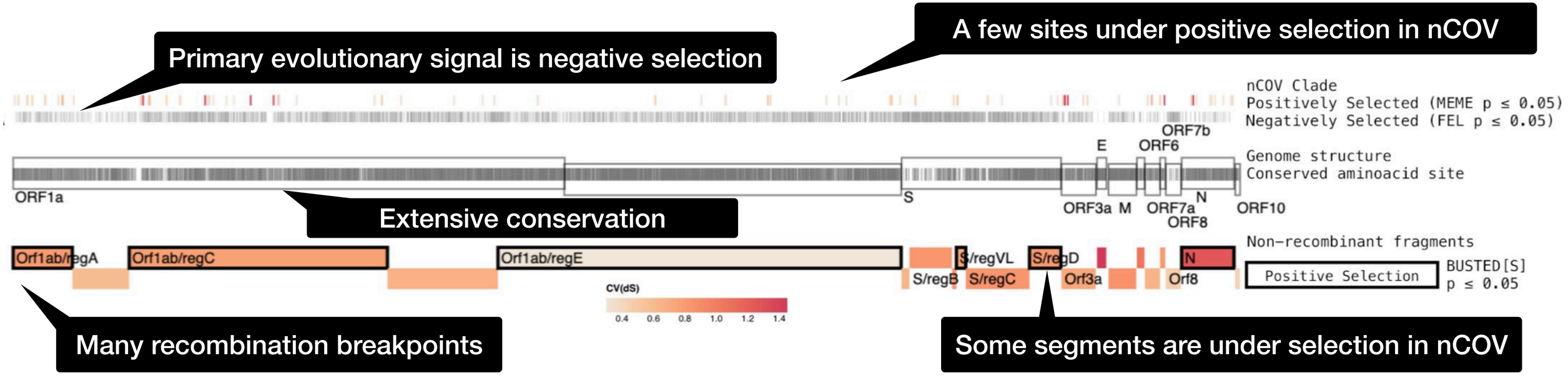
- Found evidence of extensive recombination history in the nCOV clade.
  - Later studies show evidence of cold/hot spots as well *Lytras et al 2021 BioRxiv*
- Compelling argument for CG depletion as an adaptive trait.



December 2019 | November 2020 | May 2021

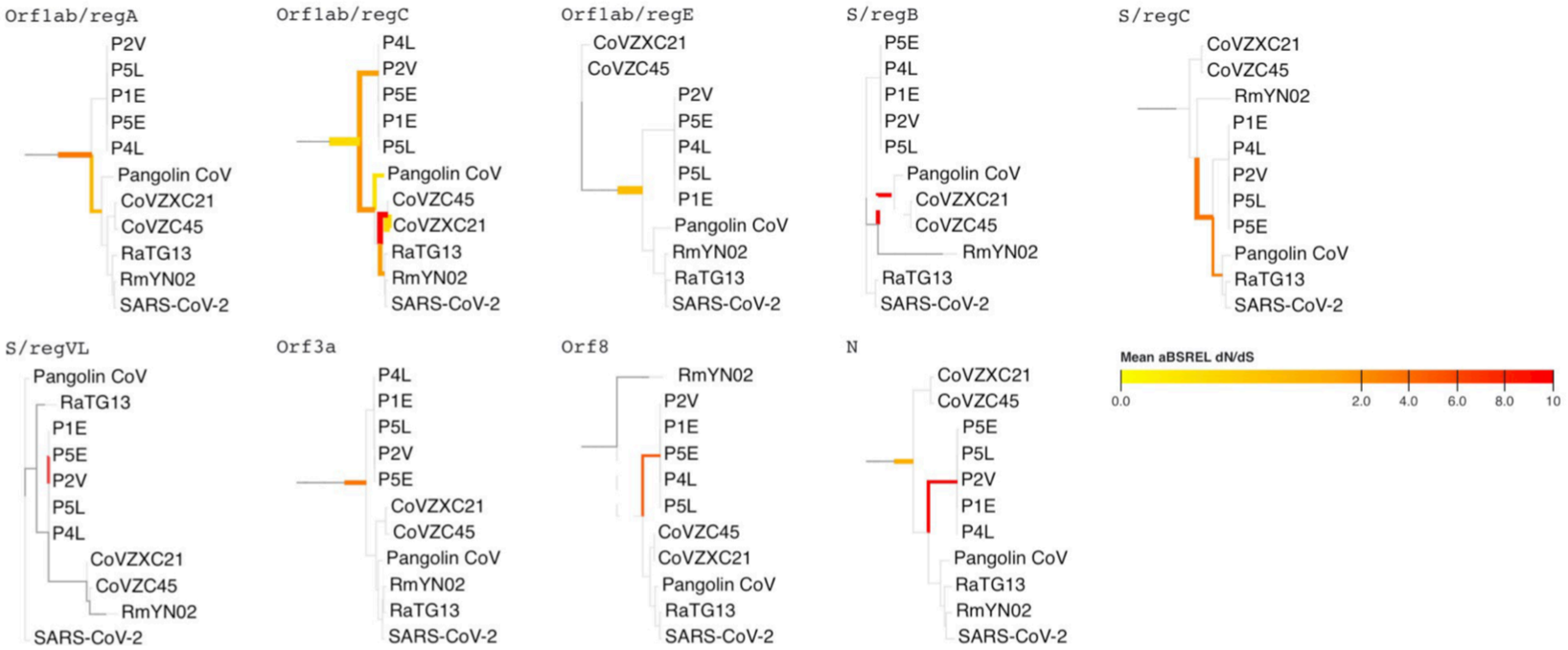
Zoonosis/early adaptation

- Found evidence of extensive recombination history in the nCOV clade.
  - Later studies show evidence of cold/hot spots as well *Lytras et al 2021 BioRxiv*
- Compelling argument for CG depletion as an adaptive trait.



December 2019 | November 2020 | May 2021

Zoonosis/early adaptation

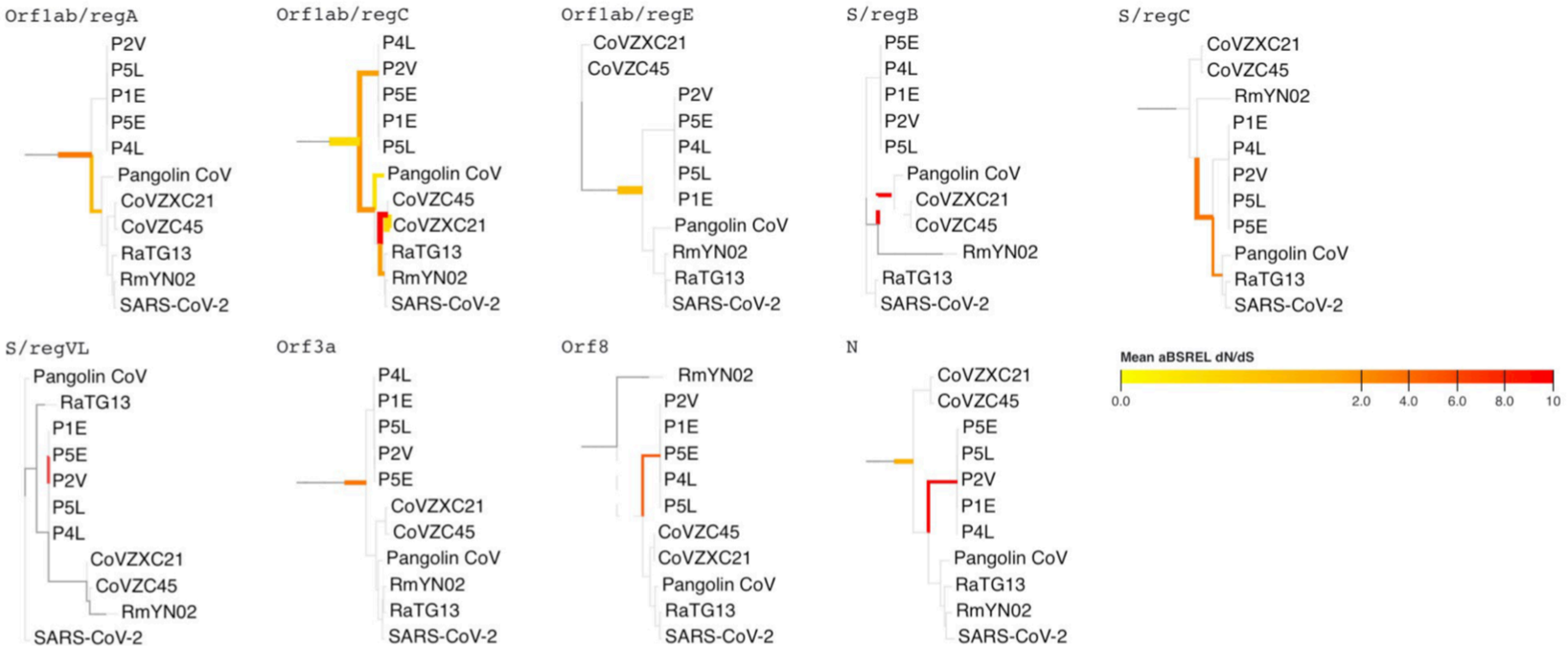


MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.

December 2019 | November 2020 | May 2021

Zoonosis/early adaptation

- Evidence of selective pressure in the nCOV clade but **not** on the SARS-CoV-2 lineage

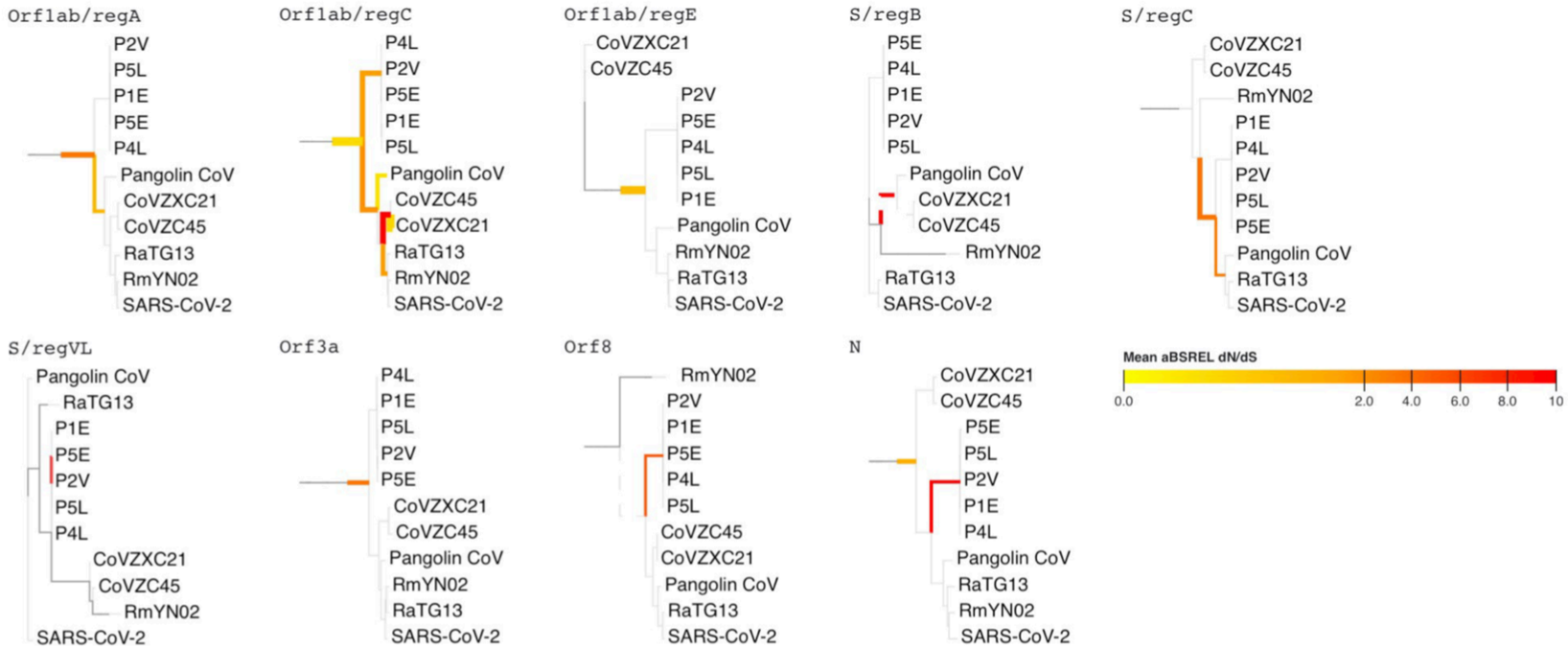


MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.

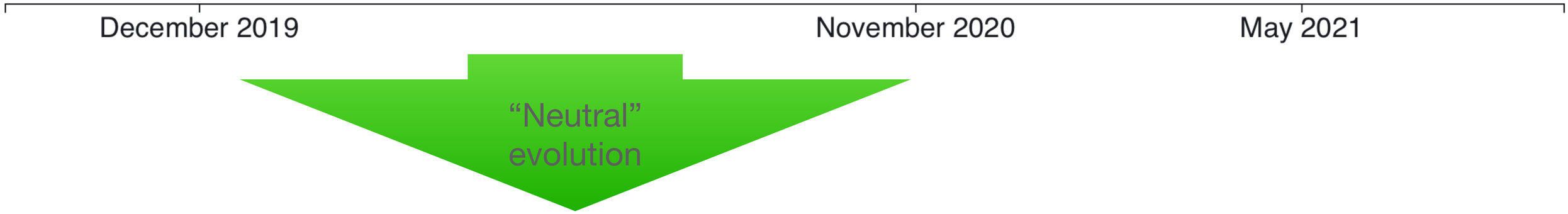
December 2019 | November 2020 | May 2021

Zoonosis/early adaptation

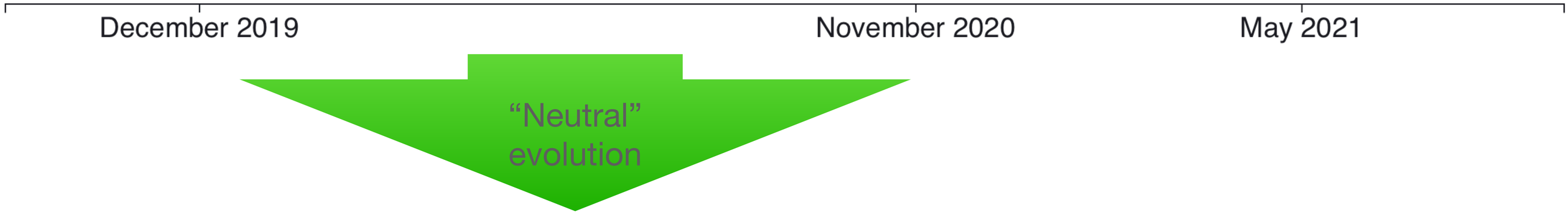
- Evidence of selective pressure in the nCOV clade but **not** on the SARS-CoV-2 lineage
- SARS-CoV-2 appears to belong to a generalist virus clade, adept at host switching and not requiring significant **immediate** adaptation during zoonosis.



MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.

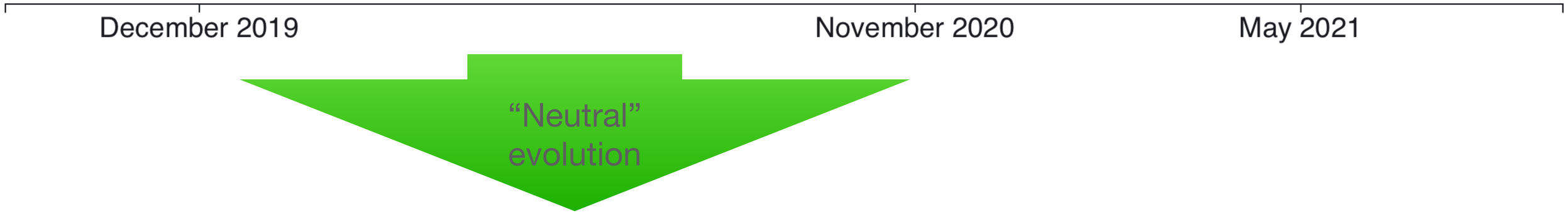


MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.



- Up to Oct 2020, the evolution of SARS-CoV-2 in humans follows a mostly neutral pattern

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.



- Up to Oct 2020, the evolution of SARS-CoV-2 in humans follows a mostly neutral pattern
- Naive hosts=>lack of selective pressure

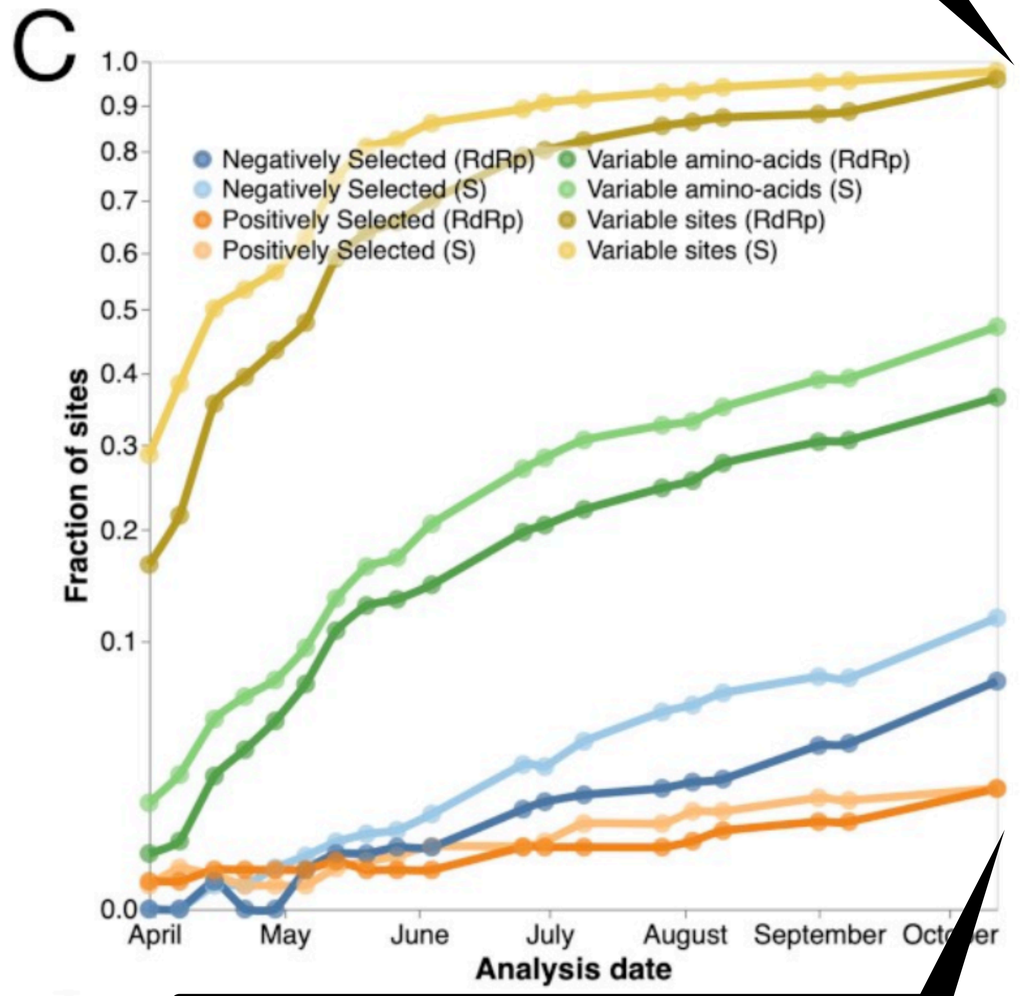
MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.

December 2019 — November 2020 — May 2021

"Neutral" evolution

- Up to Oct 2020, the evolution of SARS-CoV-2 in humans follows a mostly neutral pattern
- Naive hosts=>lack of selective pressure
- Relative lack of meaningful diversity

Lots of "raw" variation



Little "informative" variation

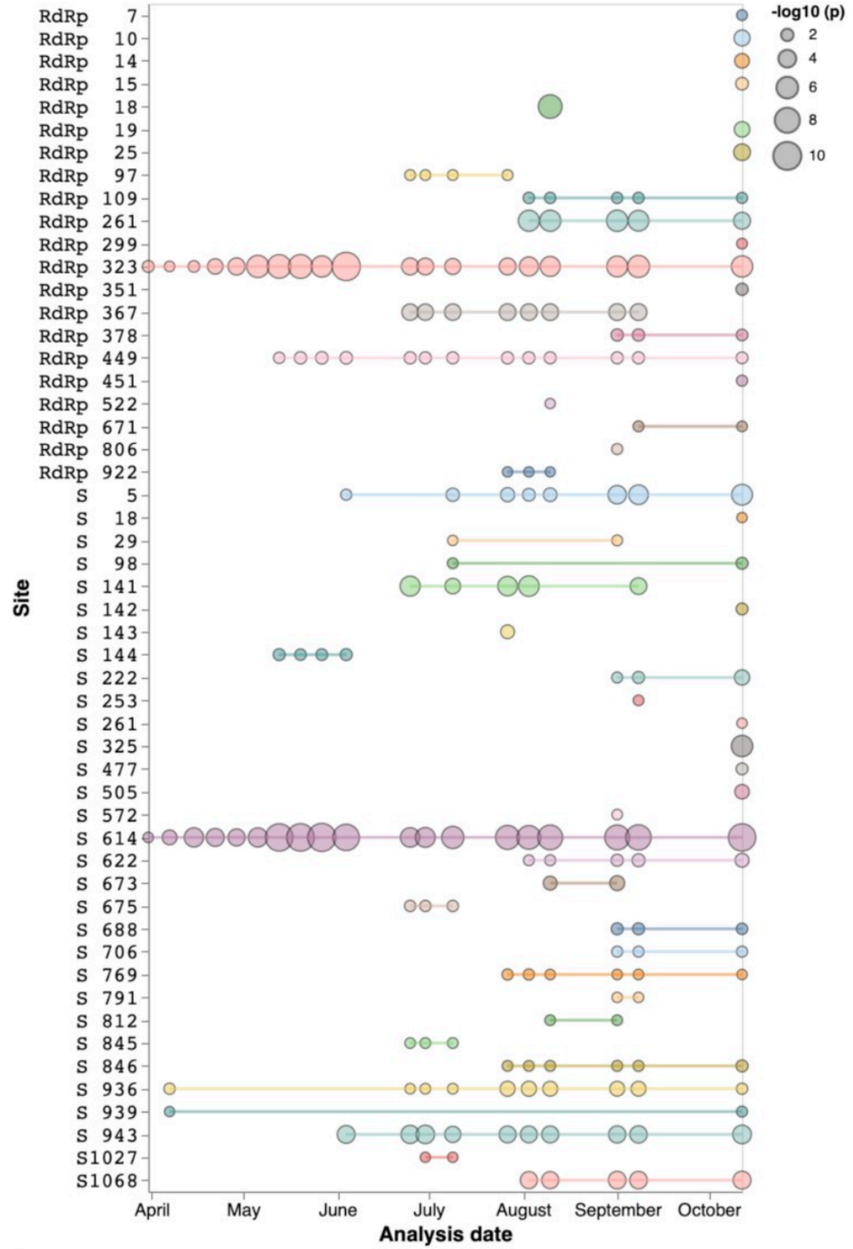
MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.

December 2019

November 2020

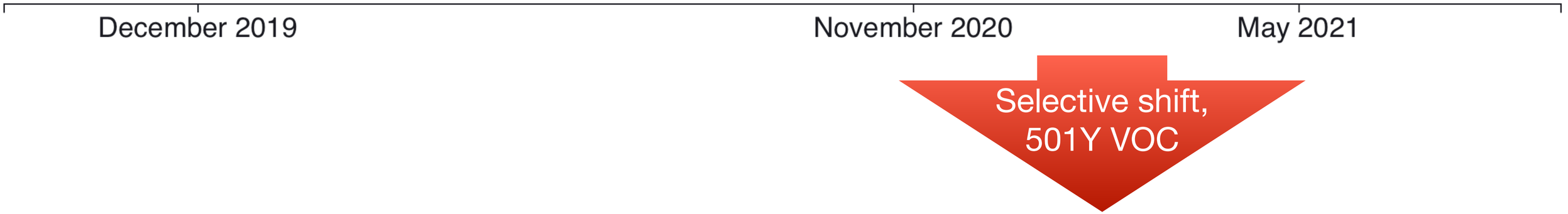
May 2021

"Neutral" evolution

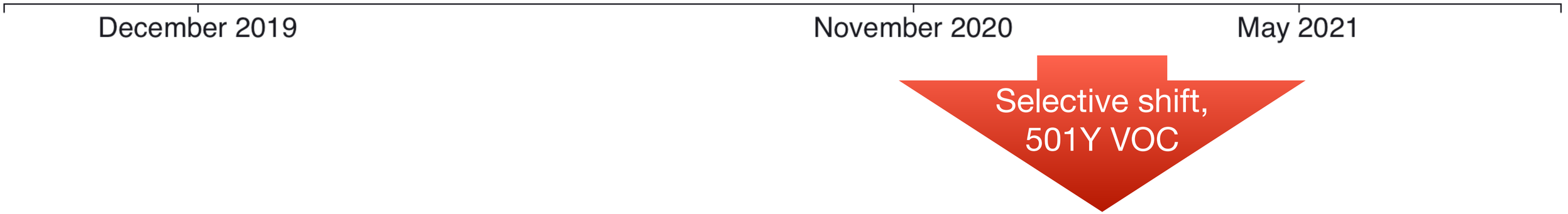


- Up to Oct 2020, the evolution of SARS-CoV-2 in humans follows a mostly neutral pattern
- Naive hosts=>lack of selective pressure
- Relative lack of meaningful diversity
- Only a few sites exhibit signal for positive selection (~150,000 GISAID sequences)

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3): e3001115.



Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages



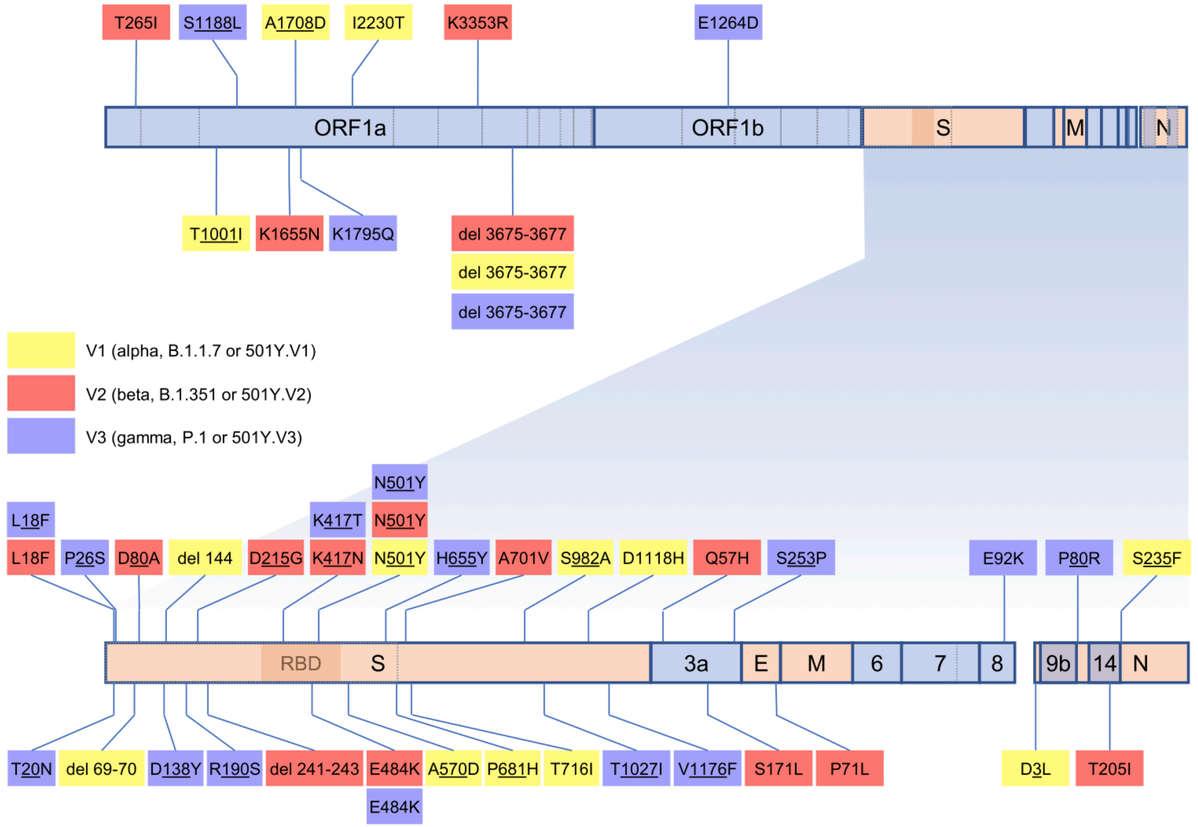
- Three **independent** emergences of N501Y lineages (alpha, beta, gamma).

Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

Selective shift, 501Y VOC

- Three **independent** emergences of N501Y lineages (alpha, beta, gamma).
- Striking patterns of shared changes.

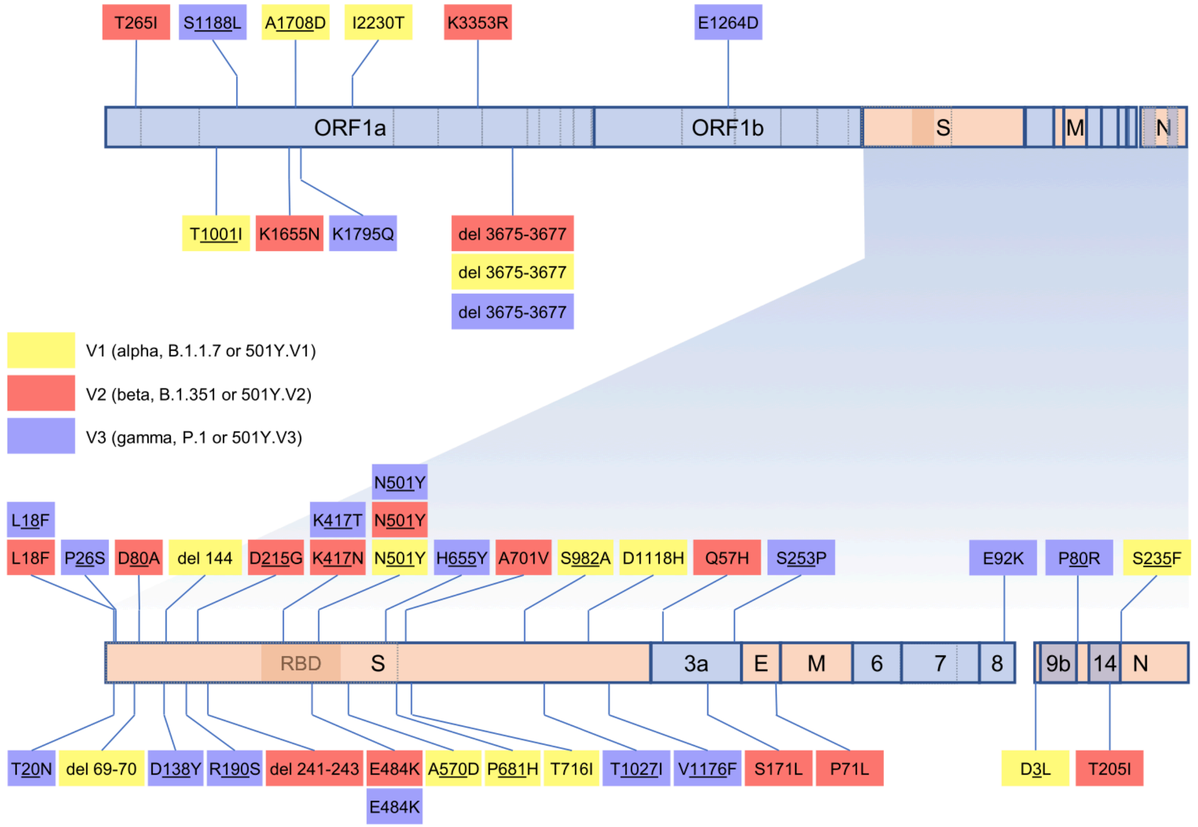


Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

Selective shift, 501Y VOC

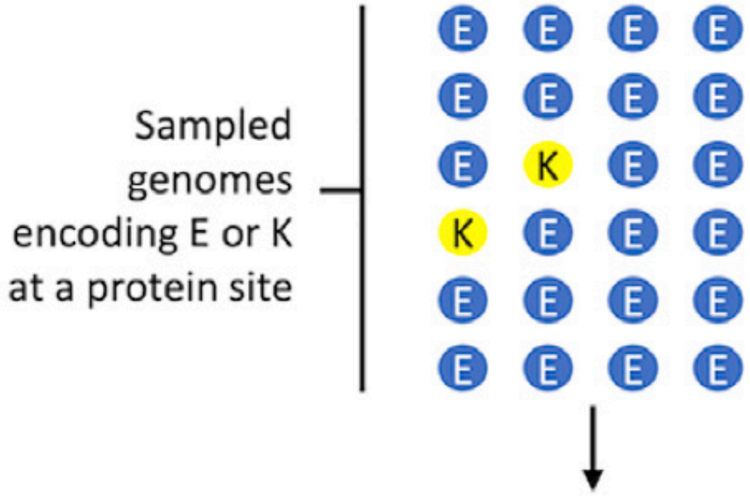
- Three **independent** emergences of N501Y lineages (alpha, beta, gamma).
- Striking patterns of shared changes.
- Set out to perform detailed analyses of >2M GISAID sequences to identify where and how convergent evolution may be operating.



Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 November 2020 May 2021

Selective shift,
501Y VOC

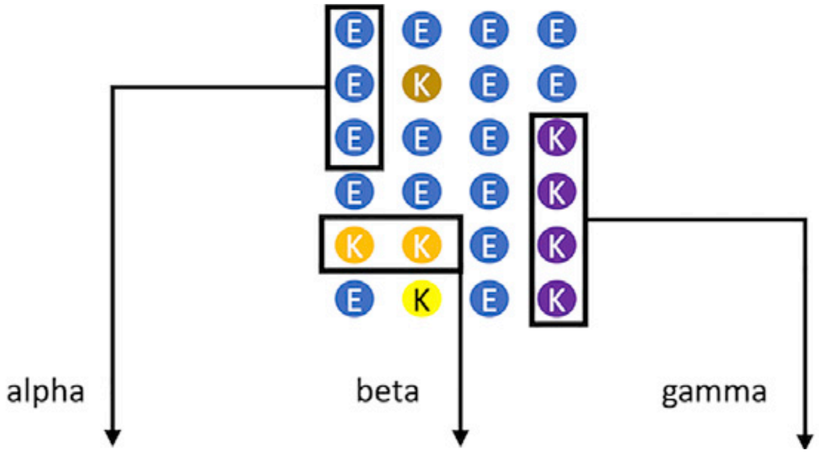


Pre-November 2020:
The evolutionary signal in the codon encoding this protein site fails to trigger a positive selection signal

Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

Selective shift,
501Y VOC



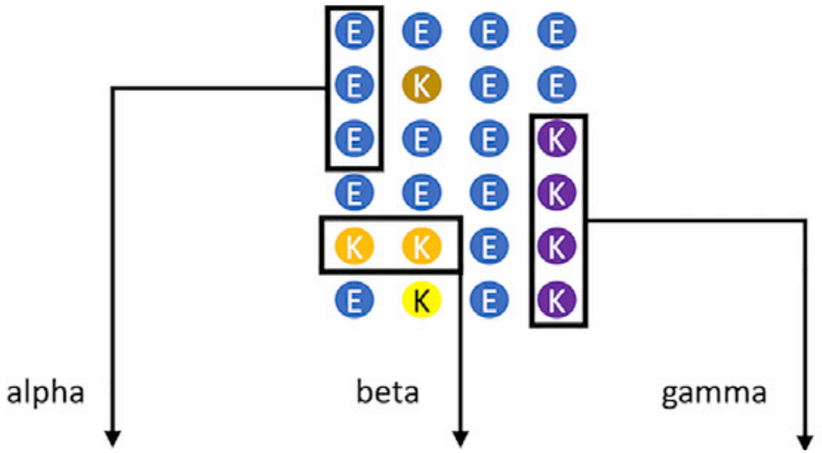
November 2020:
An excess of independent mutations (note different colours) that change the encoded amino acid at this site triggers a positive selection signal

December 2020
Discovery of the alpha, beta and gamma lineages. It is apparent that K is favoured over E at this site

Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

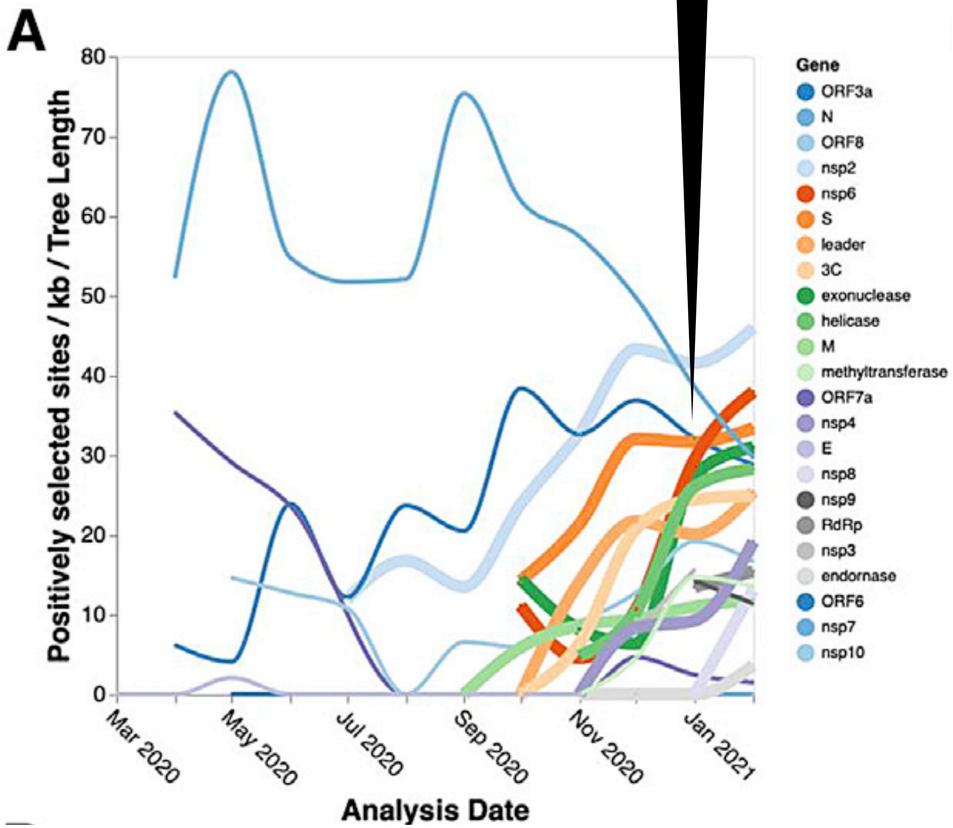
Selective shift, 501Y VOC



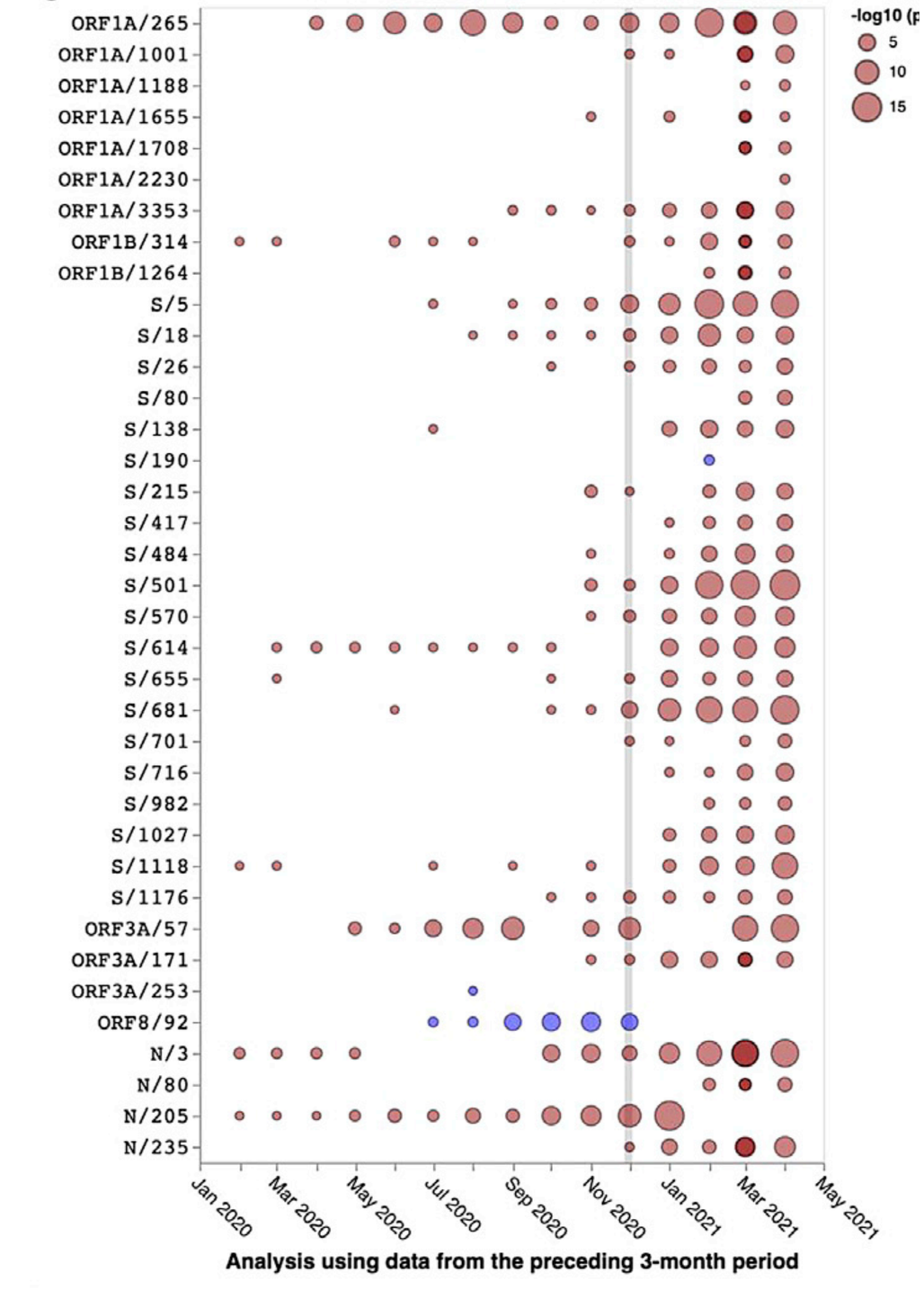
November 2020:
An excess of independent mutations (note different colours) that change the encoded amino acid at this site triggers a positive selection signal

December 2020
Discovery of the alpha, beta and gamma lineages. It is apparent that K is favoured over E at this site

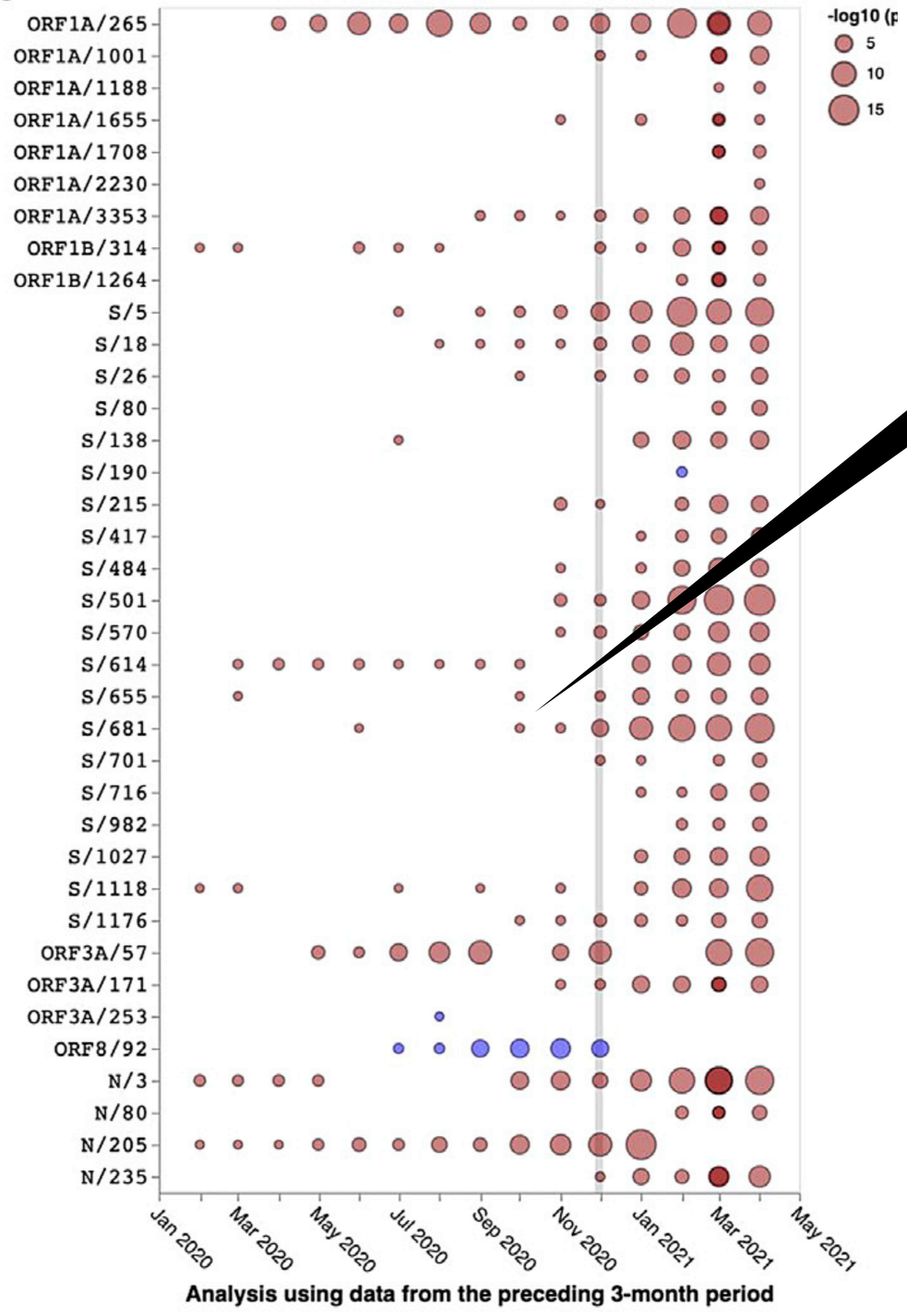
Ramp-up in selective pressure



Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages



Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

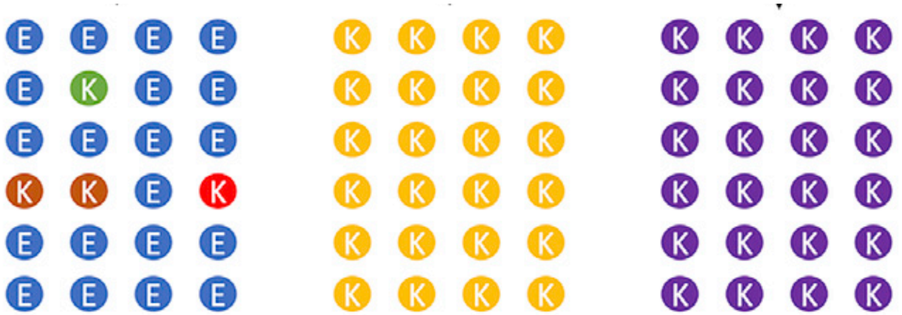


Many “clade-defining” mutations were detectably selected BEFORE they became clade defining

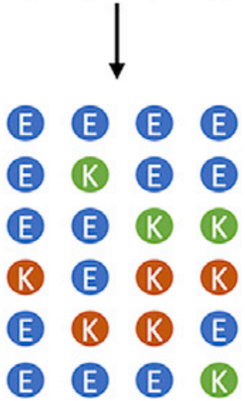
Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

Selective shift, 501Y VOC



March 2021
E to K associated positive selection signal detectable even when only considering alpha sequences



June 2021
K more than doubles in frequency between March and June 2021, reiterating that K provides a fitness advantage: K at this site is added to the 501Y meta-signature along with various mutations at 34 other similarly evolving genome sites

Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

Selective shift, 501Y VOC

Used data up to February to define 34 convergent sites between three N501Y lineages

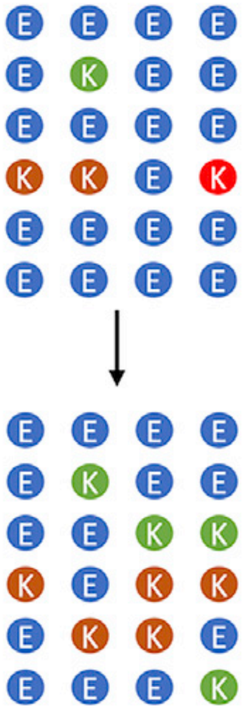
March 2021
E to K associated positive selection signal detectable even when only considering alpha sequences

June 2021
K more than doubles in frequency between March and June 2021, reiterating that K provides a fitness advantage: K at this site is added to the 501Y meta-signature along with various mutations at 34 other similarly evolving genome sites

Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 November 2020 May 2021

Us



Evolutionary Probability

0.000001 0.0001 0.01 0.05 0.5

Nucleotide	Site	V1	V2	V3	V1 subs to	V2 subs to	V3 subs to
21574	S/5	✕	✕	✕	F	F	F
21613	S/18	✕	~✕	~✕	F	F	F
21619	S/20			~+		T I	N
21637	S/26	✕		~✕	S	R	S
21799	S/80	✕	~✕		A	A D S	
21853	S/98	✕	✕		F	F	S
21973	S/138	✕	✕	~✕	H Y	Y	Y D
22129	S/190			~+			S
22204	S/215	✕	~✕		G	G V D H	
22810	S/417		~✕	~✕		N K	T K
23011	S/484	✕	~✕	~✕	K Q	K E	K
23062	S/501	~✕	~✕	~✕	Y N E	Y N	Y N

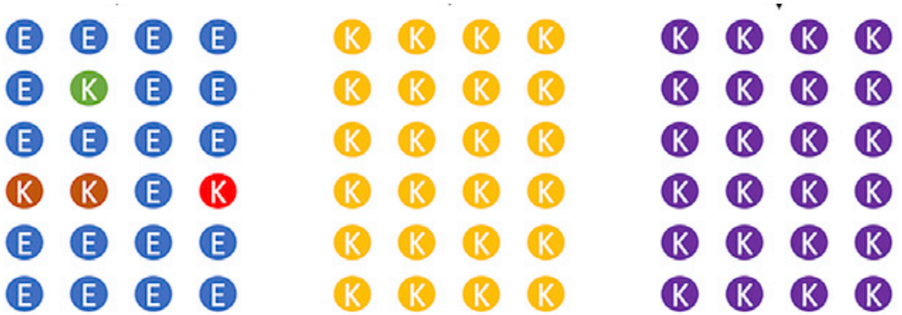
Selection legend

1. ~ : Signtature mutation
2. + : MEME support
3. ✕ : convergent subsitutions to another lineage (no MEME support)
4. ✕ : convergent subsitutions to another lineage AND MEME support

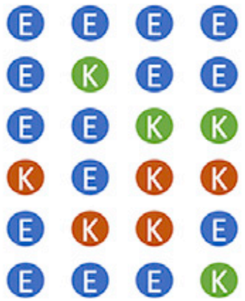
Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

Selective shift, 501Y VOC



March 2021
E to K associated positive selection signal detectable even when only considering alpha sequences



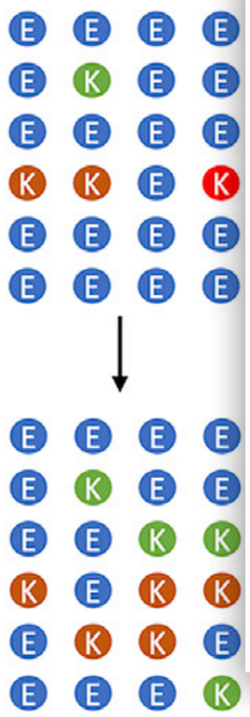
June 2021
K more than doubles in frequency between March and June 2021, reiterating that K provides a fitness advantage: K at this site is added to the 501Y meta-signature along with various mutations at 34 other similarly evolving genome sites

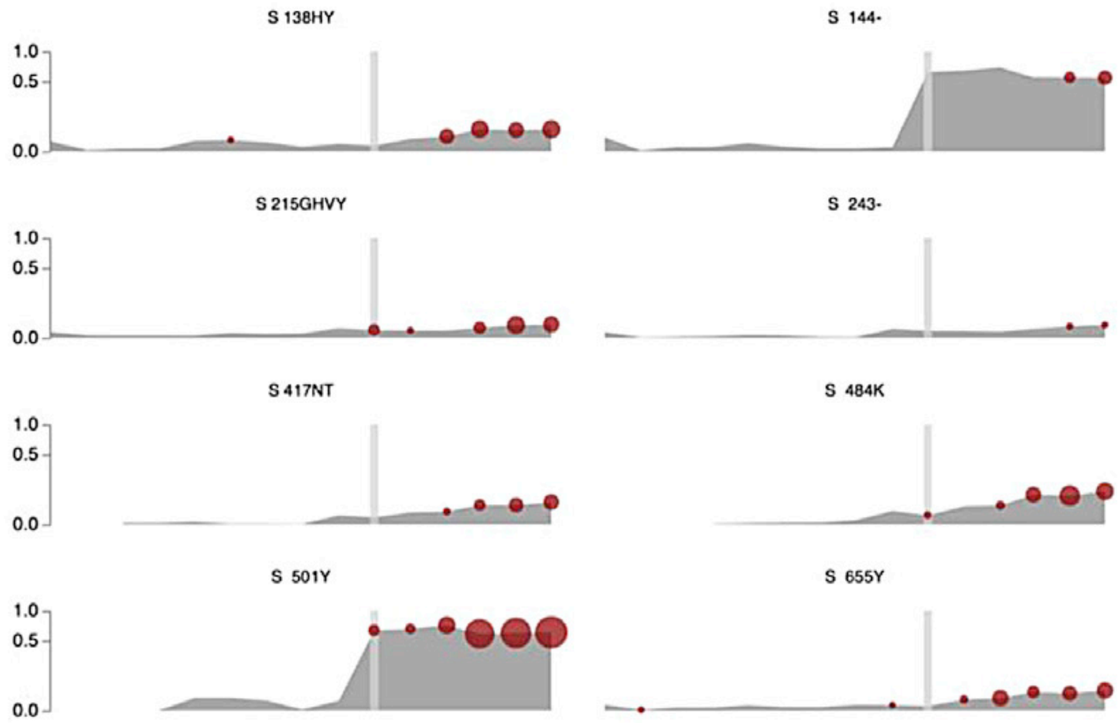
Used data up to June to validate convergent sites

Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

Selective shift, 501Y VOC





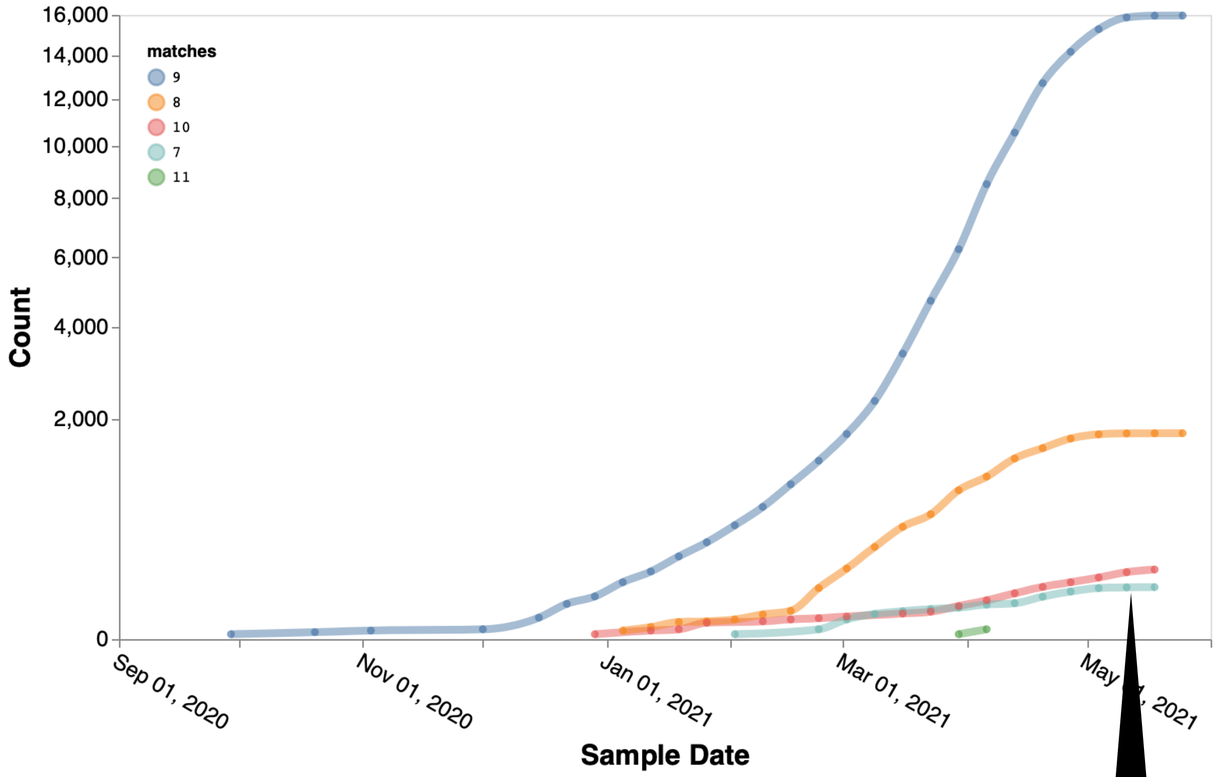
"In total, 20/47 of the analyzed convergent mutations that were suggested by our global and lineage-specific positive selection analyses to have contributed to the fitness of 501Y lineage viruses prior to March 2021 more than doubled in frequency in at least one of the lineages between 15 March and June 1, 2021"

Used data up to June to validate convergent sites

Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

Selective shift,
501Y VOC



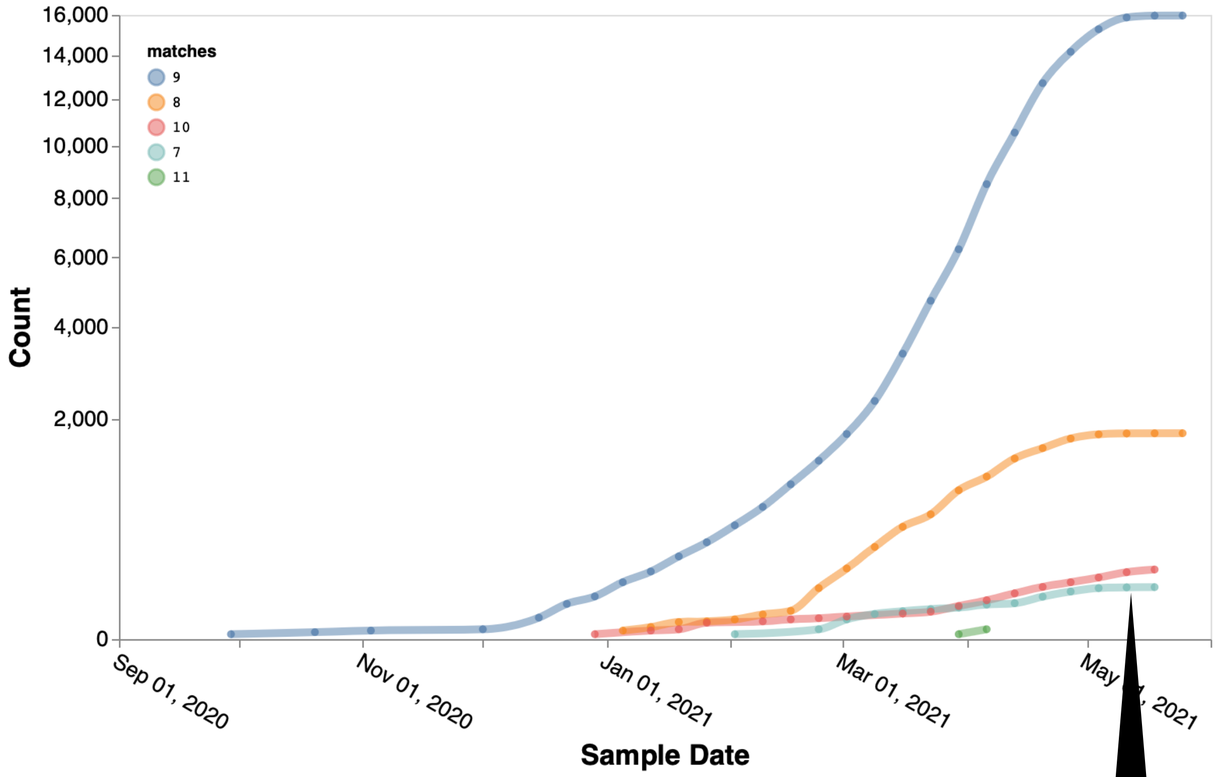
Sequences acquiring additional mutations in Spike (V3)

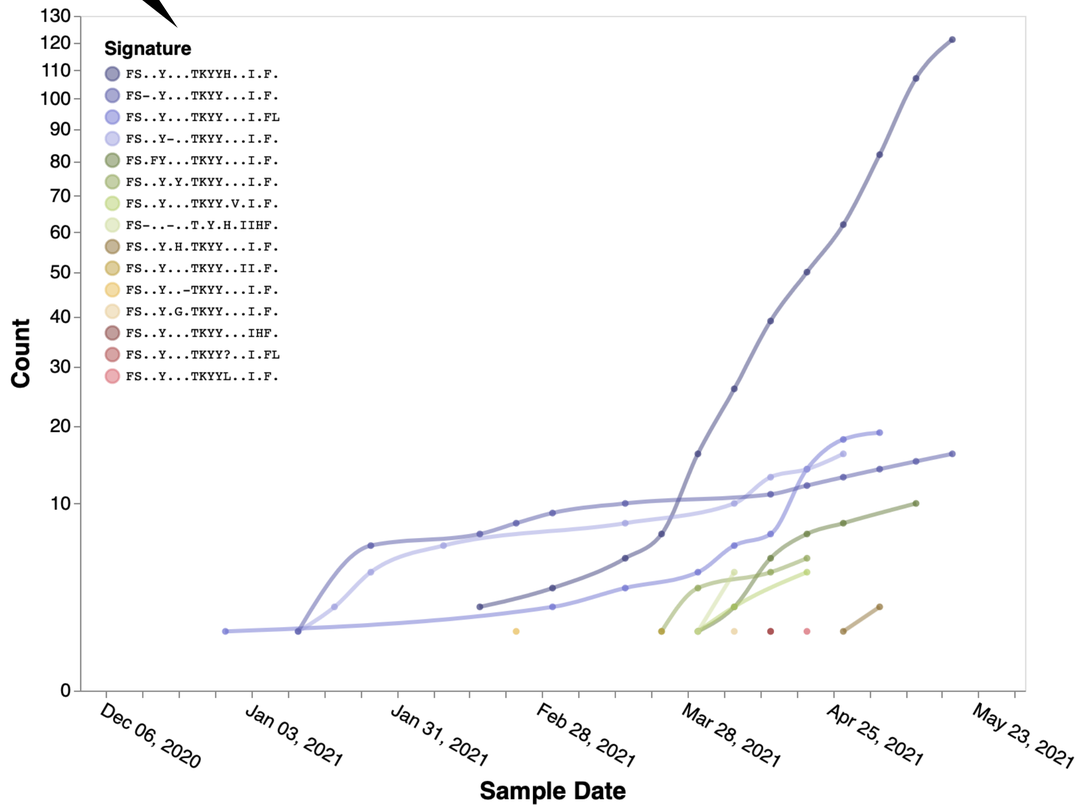
Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019 | November 2020 | May 2021

Selective shift, 501Y VOC

Spike signatures in “most converged” sequences





Sequences acquiring additional mutations in Spike (V3)

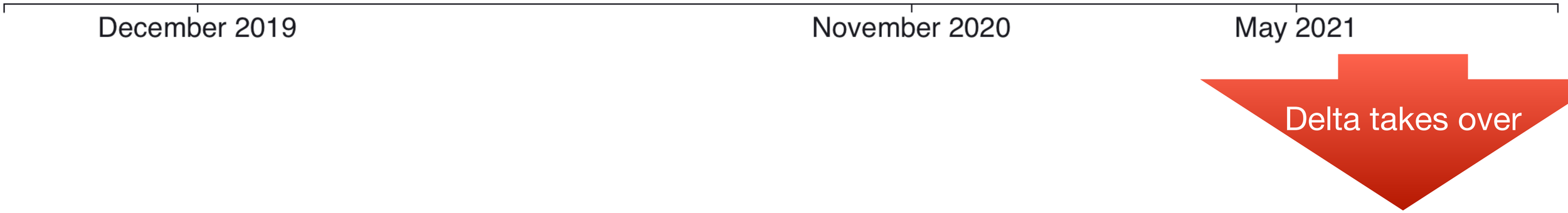
Martin et al., 2021, Cell 184, 5189–5200 The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages

December 2019

November 2020

May 2021

Delta takes over



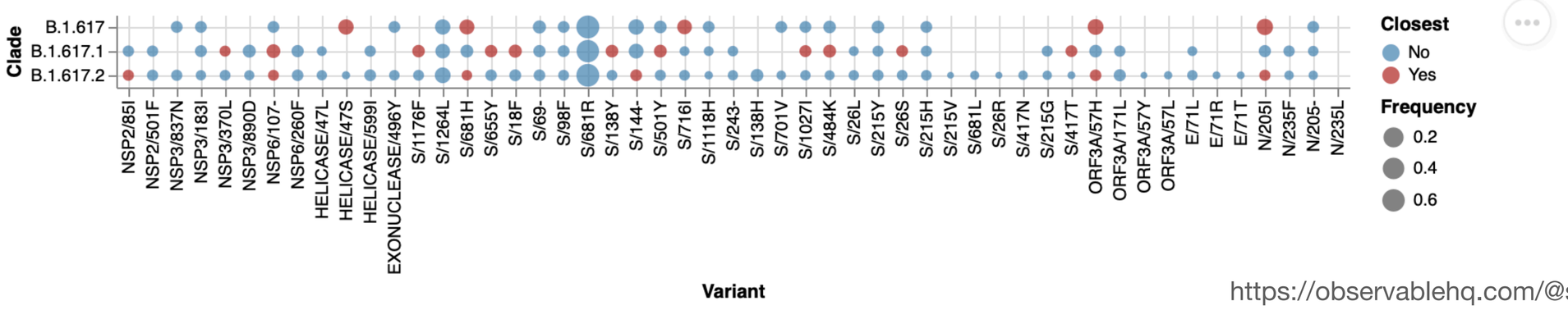
- The N501Y lineages have been supplanted by the Delta lineage in the Summer of 2021.
- These “S/452” lineages appear to be on a different fitness peak compared to “S/501” lineages.
- Is our metasignature “dead”?

December 2019 | November 2020 | May 2021

Delta takes over

B.1.617 3 results

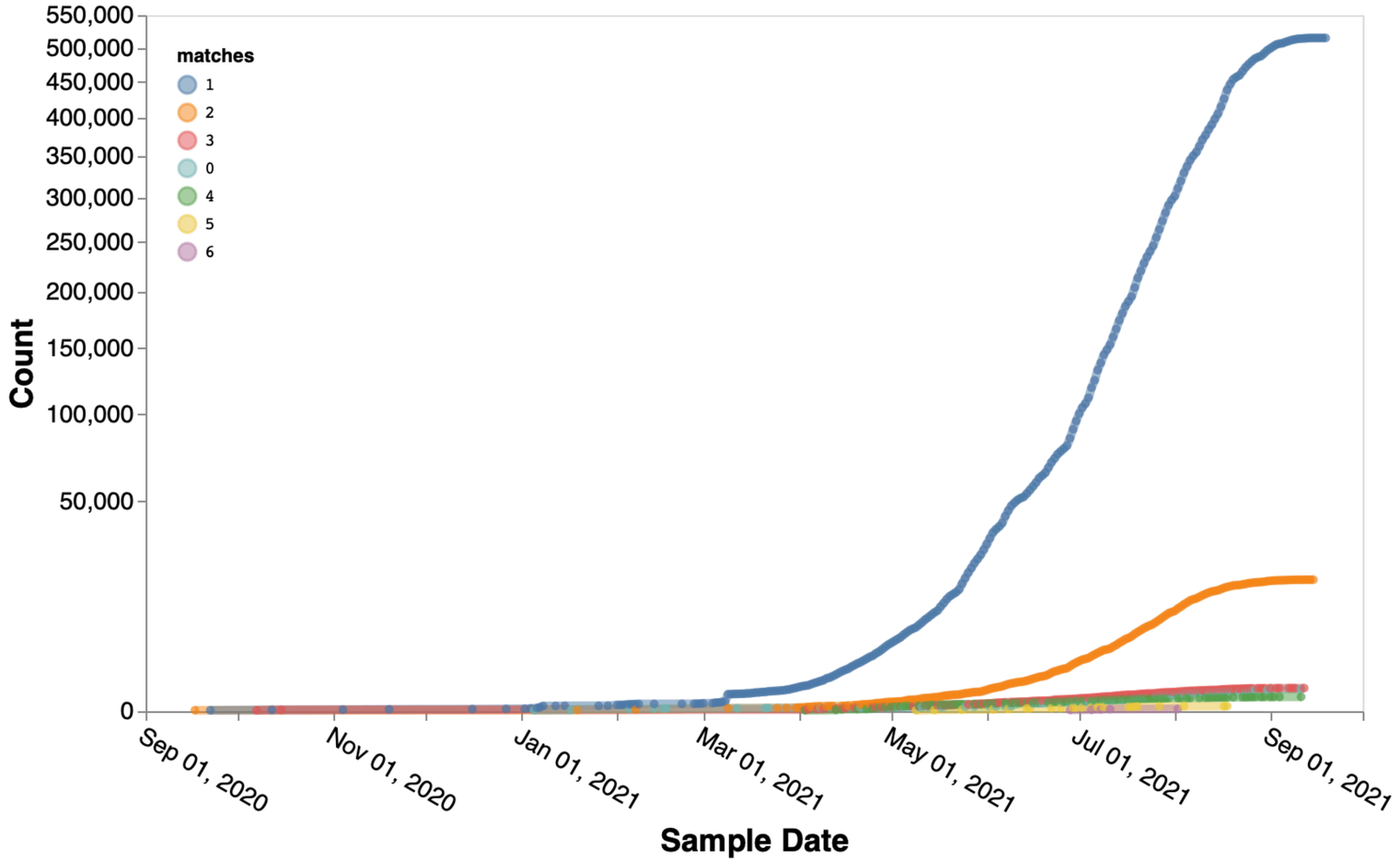
Clade	N	Counts	Closest (≥3 seqs)	Found in	Matches
B.1.617	1,020		5	6	HELICASE/47S, S/681H, S/716I, ORF3A/57H, N/205I
B.1.617.1	6,070		11	3	NSP3/370L, NSP6/107-, S/18F, S/26S, S/138Y, S/417T, S/484K, S/501Y, S/655Y, S/1027I, S/1176F
B.1.617.2	748,130		6	46	NSP2/85I, NSP6/107-, S/144-, S/681H, ORF3A/57H, N/205I



https://observablehq.com/@spond/n501y-metasignature

December 2019 | November 2020 | May 2021

Delta takes over

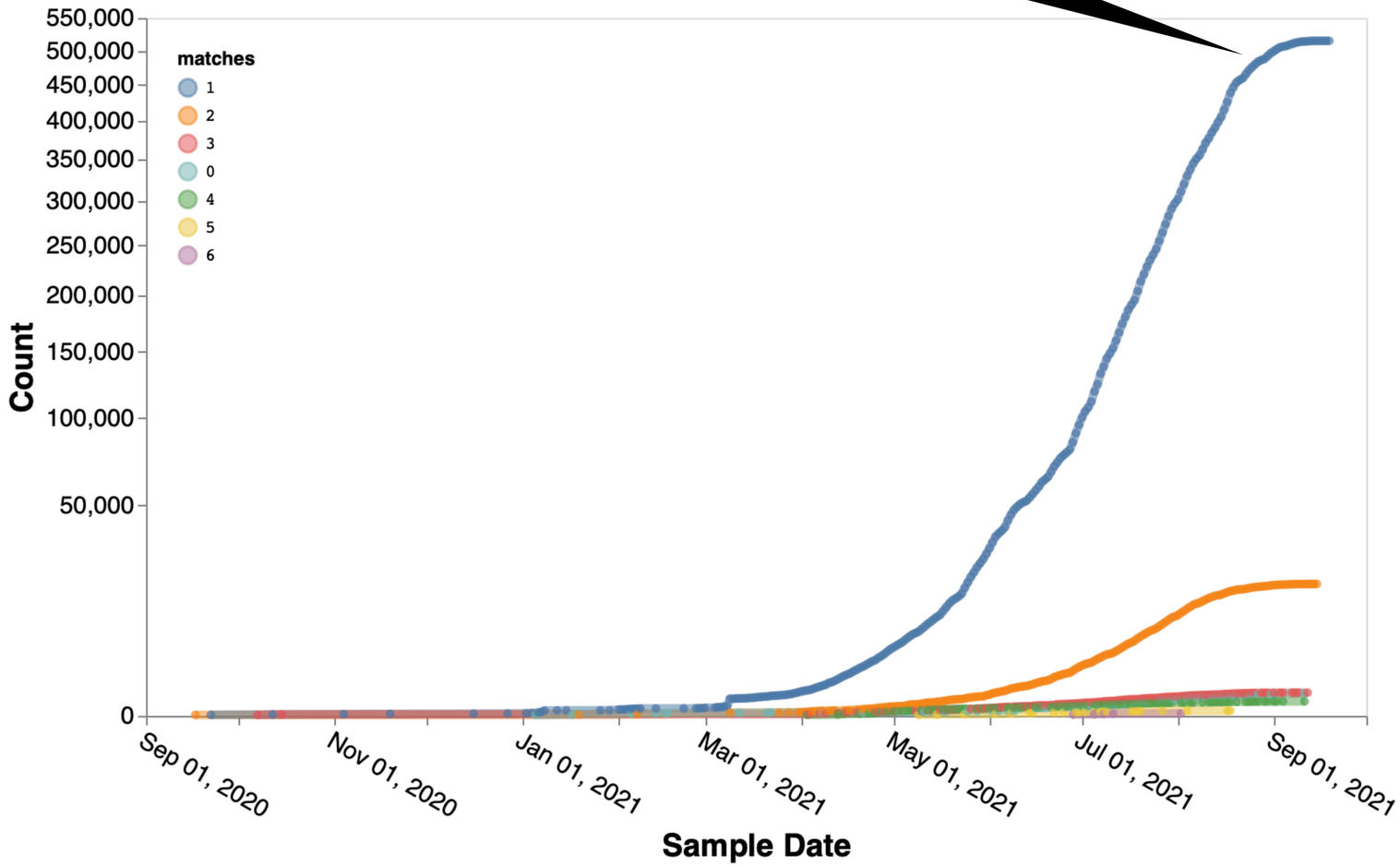


https://observablehq.com/@spond/spike-trends

December 2019 | November 2020 | May 2021

Most sequences in B.1.617.2 have a single match to the meta-signature in Spike (S/681R)

Delta takes over

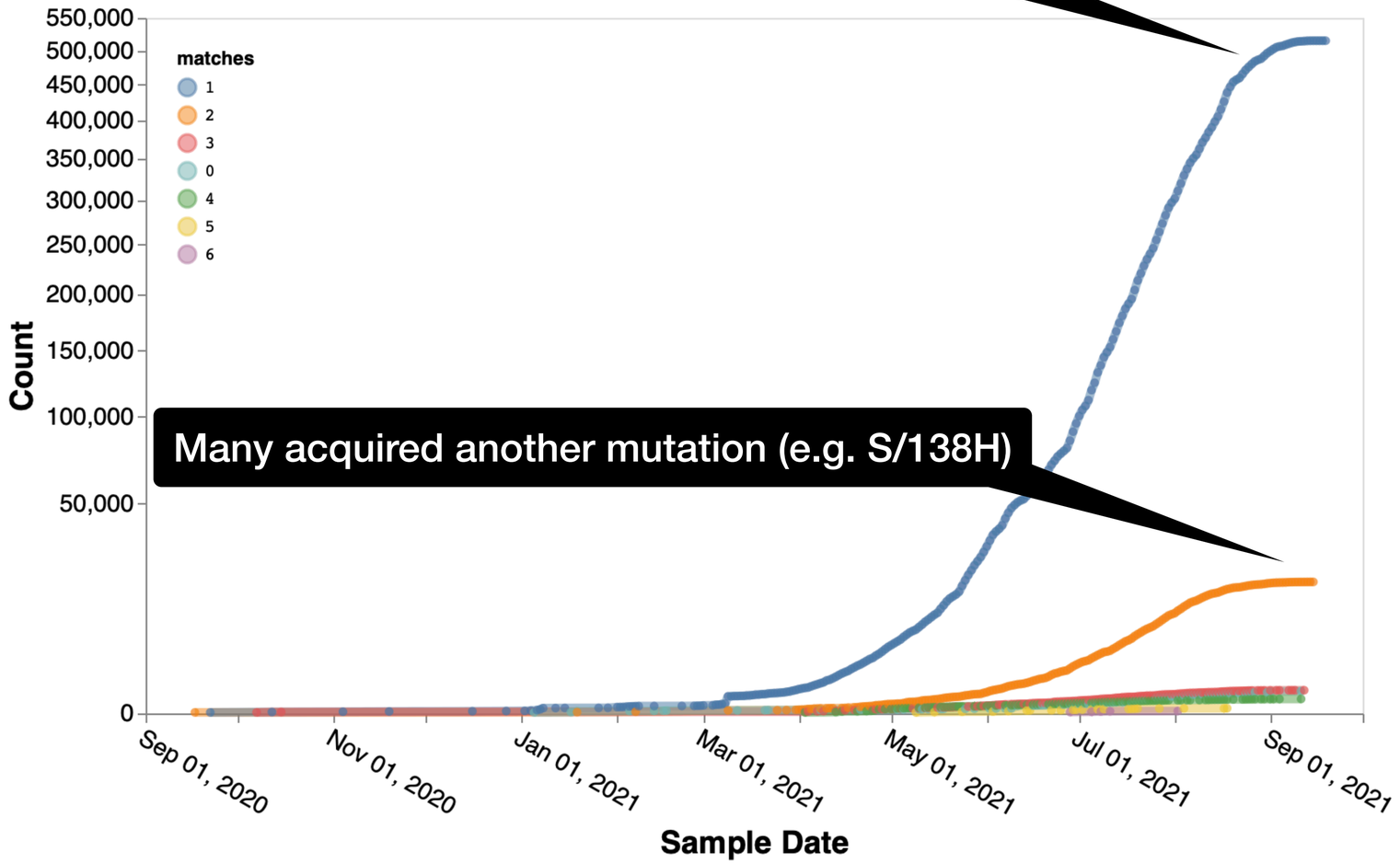


https://observablehq.com/@spond/spike-trends

December 2019 | November 2020 | May 2021

Delta takes over

Most sequences in B.1.617.2 have a single match to the meta-signature in Spike (S/681R)



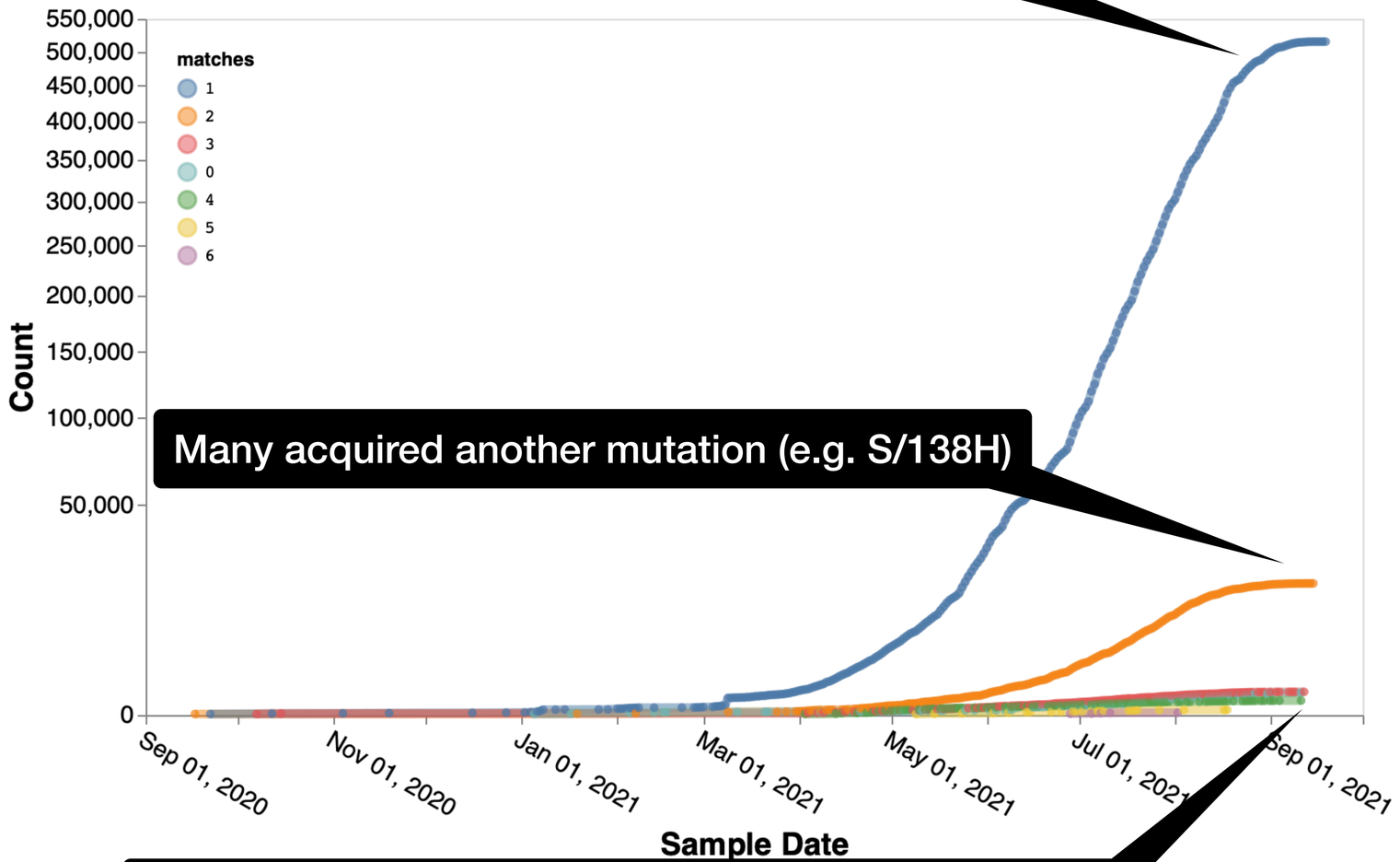
Many acquired another mutation (e.g. S/138H)

https://observablehq.com/@spond/spike-trends

December 2019 | November 2020 | May 2021

Delta takes over

Most sequences in B.1.617.2 have a single match to the meta-signature in Spike (S/681R)



Many acquired another mutation (e.g. S/138H)

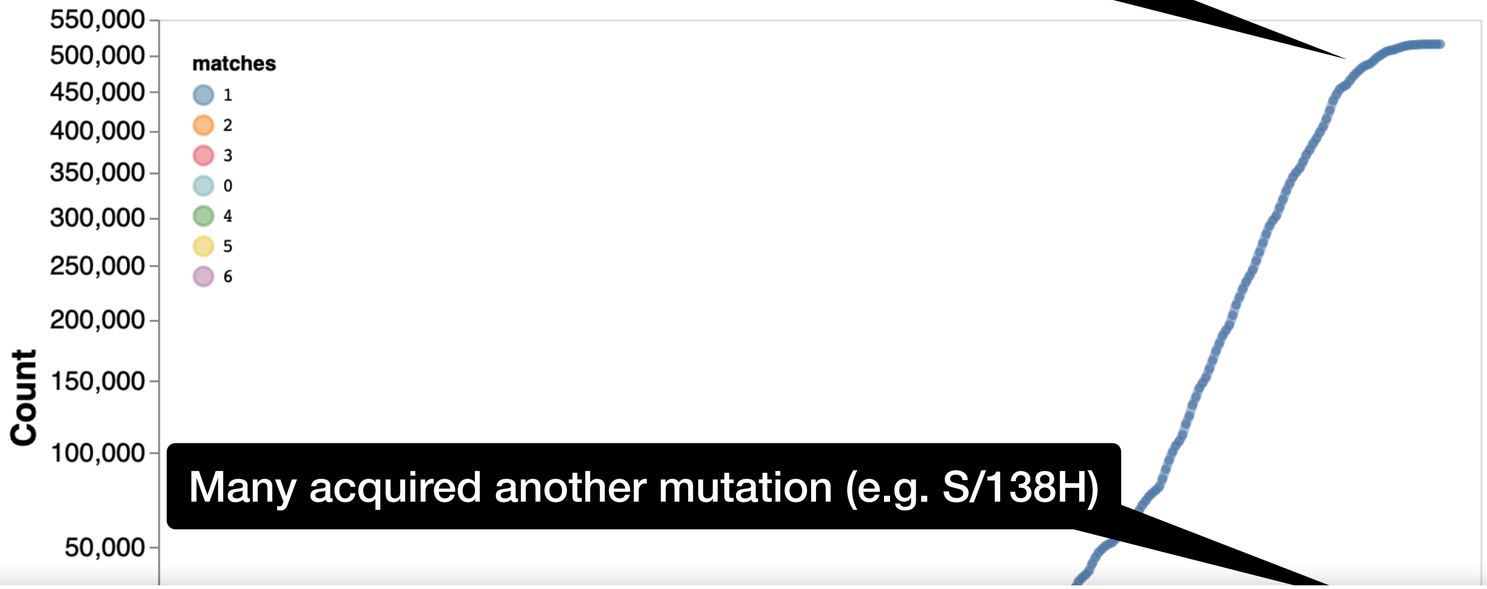
Some sequences acquired 3, 4 or 5 additional signature mutations

https://observablehq.com/@spond/spike-trends

December 2019 November 2020 May 2021

Delta takes over

Most sequences in B.1.617.2 have a single match to the meta-signature in Spike (S/681R)

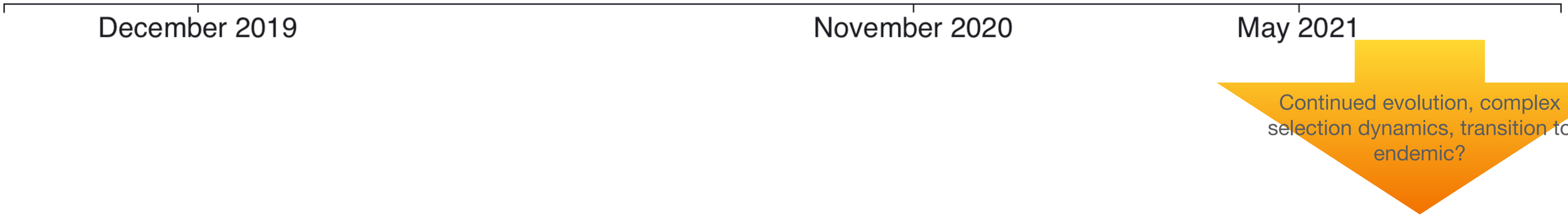


Many acquired another mutation (e.g. S/138H)

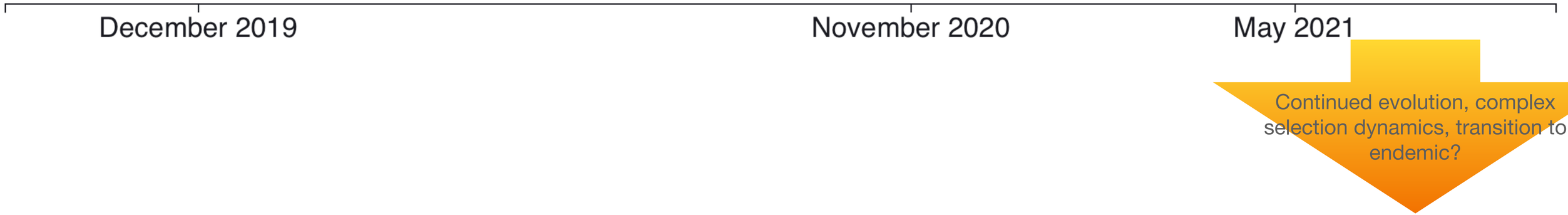
ID	Sampled	Location	Clade	Matches ↓
2485255	2021/08/02	North America, USA, Nevada	B.1.617.2/21A (Delta)	69-, 70-, 98F, 138H, 144-, 681R
2485522	2021/07/11	North America, USA, Oregon	B.1.617.2/21A (Delta)	69-, 70-, 98F, 138H, 144-, 681R
2249409	2021/07/08	Europe, Greece, Attica	B.1.617.2/21A (Delta)	69-, 70-, 501Y, 681R, 716I, 1118H
1847594	2021/07/05	Europe, Sweden, Sodermanland	B.1.617.2/21A (Delta)	69-, 70-, 144-, 681H, 716I, 1118H
1715407	2021/06/28	North America, Mexico, Quintana Roo	B.1.617.2/21A (Delta)	18F, 26S, 681R, 1027I, 1176F, 1264L

Some sequences acquired 3, 4 or 5 additional signature mutations

https://observablehq.com/@spond/spike-trends



- Can comparative analyses help prioritize genomic variants (and locations of variation) that are important in some sense?
  - Likely to increase in frequency in the near future
  - Involved in adaptation
  - Show “unexpected” evolutionary patterns.



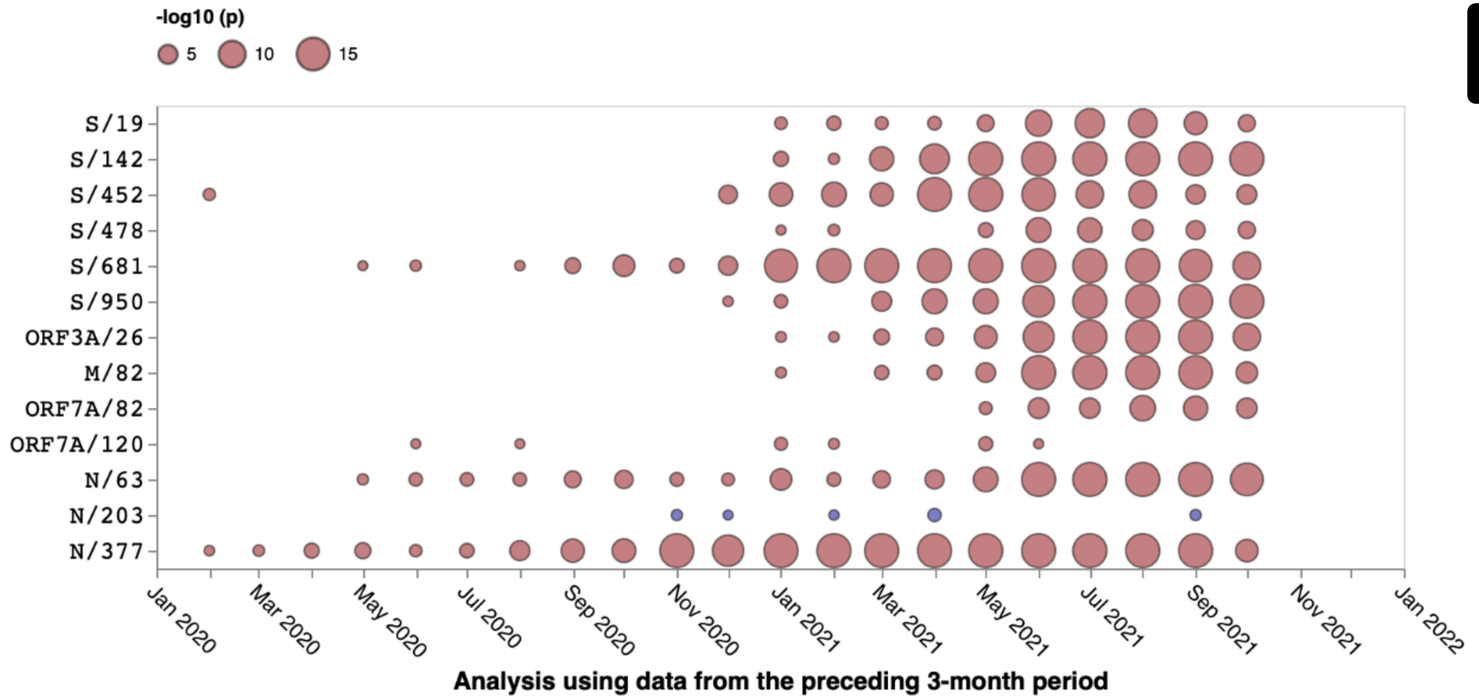
- Selection analyses (global or clade level) seem to have the ability to pick up selective signals **before** they sweep up in frequencies, become VOI/VOC

https://observablehq.com/@spond/sars-cov-2-selected-sites

December 2019 | November 2020 | May 2021

Continued evolution, complex selection dynamics, transition to endemic?

- Selection analyses (global or clade level) seem to have the ability to pick up selective signals **before** they sweep up in frequencies, become VOI/ VOC



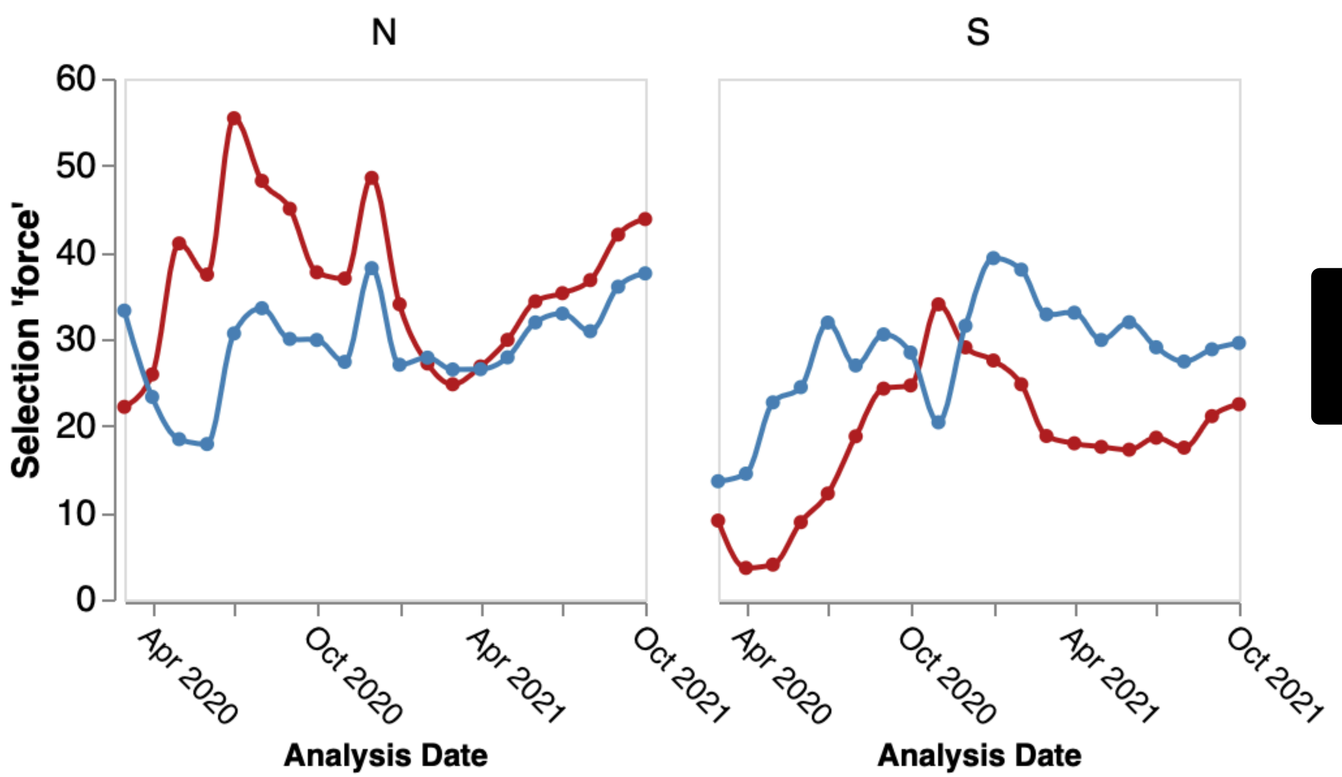
Selection at delta (B.1.617.2) sites

https://observablehq.com/@spond/sars-cov-2-selected-sites

December 2019 | November 2020 | May 2021

Continued evolution, complex selection dynamics, transition to endemic?

Figure 1. Temporal evolution of selection force, defined as the number of positively and negatively selected sites normalized by kilobase of gene length and the internal tree length (`sites/[substitutions across the tree x gene length]`); this quantity should be directly comparable between genes and time points. Genes are sorted by maximal 'force' of positive selection over all time.

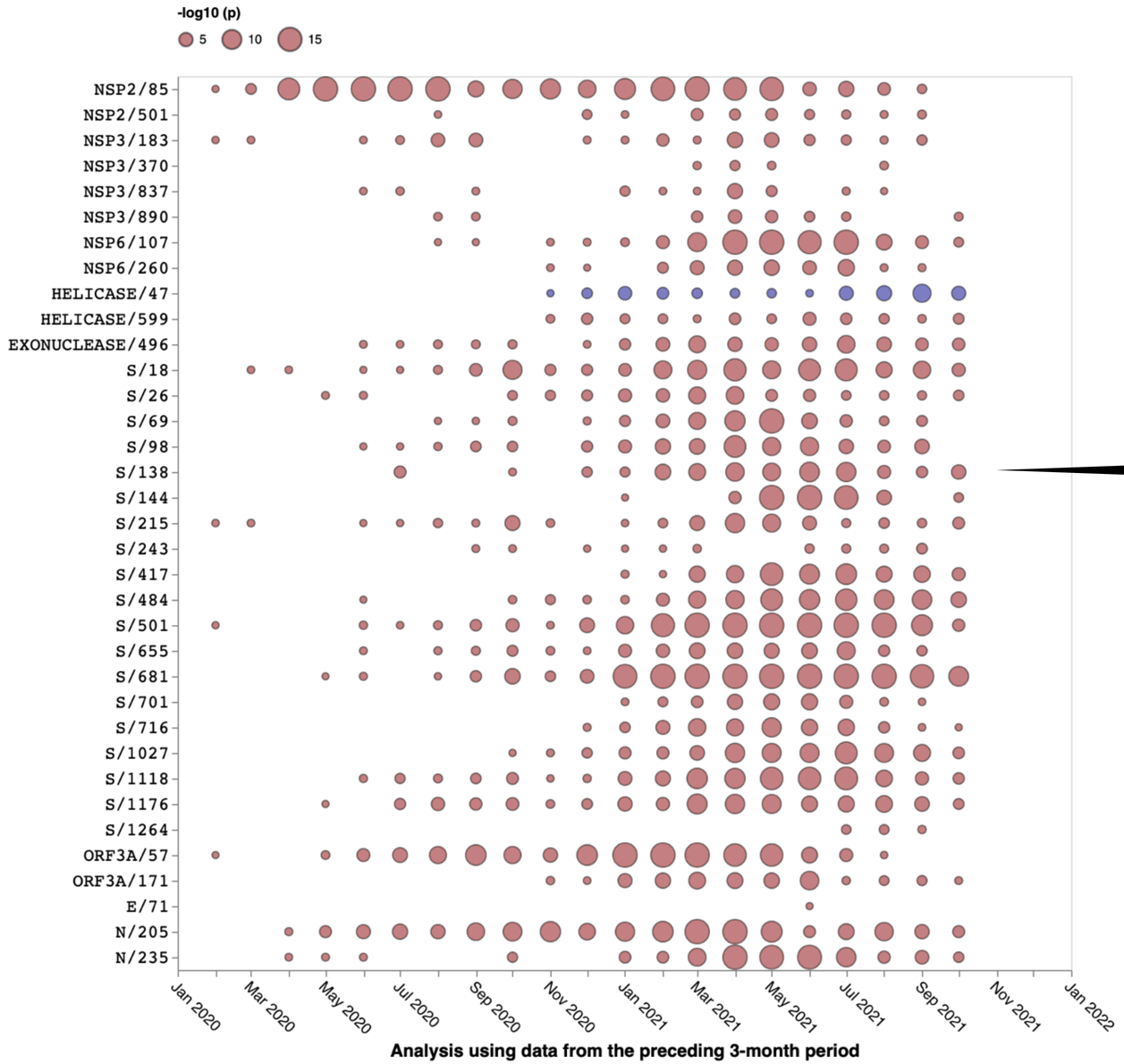


Selective force dynamics

https://observablehq.com/@spond/selection-profile

December 2019 | November 2020 | May 2021

Continued evolution, complex selection dynamics, transition to endemic?



Ongoing positive selection at a number of meta-signature sites

https://observablehq.com/@spond/sars-cov-2-selected-sites

December 2019 | November 2020 | May 2021

Continued evolution, complex selection dynamics, transition to endemic?

View history of natural selection pressures on the SARS-CoV-2 genome, through analyses of **21** three-month overlapping intervals going back to the beginning of the pandemic. The earliest intervals ends in *February 2020* and the latest - in *October 2021*.

Examine the evolution of individual sites or Download .csv data

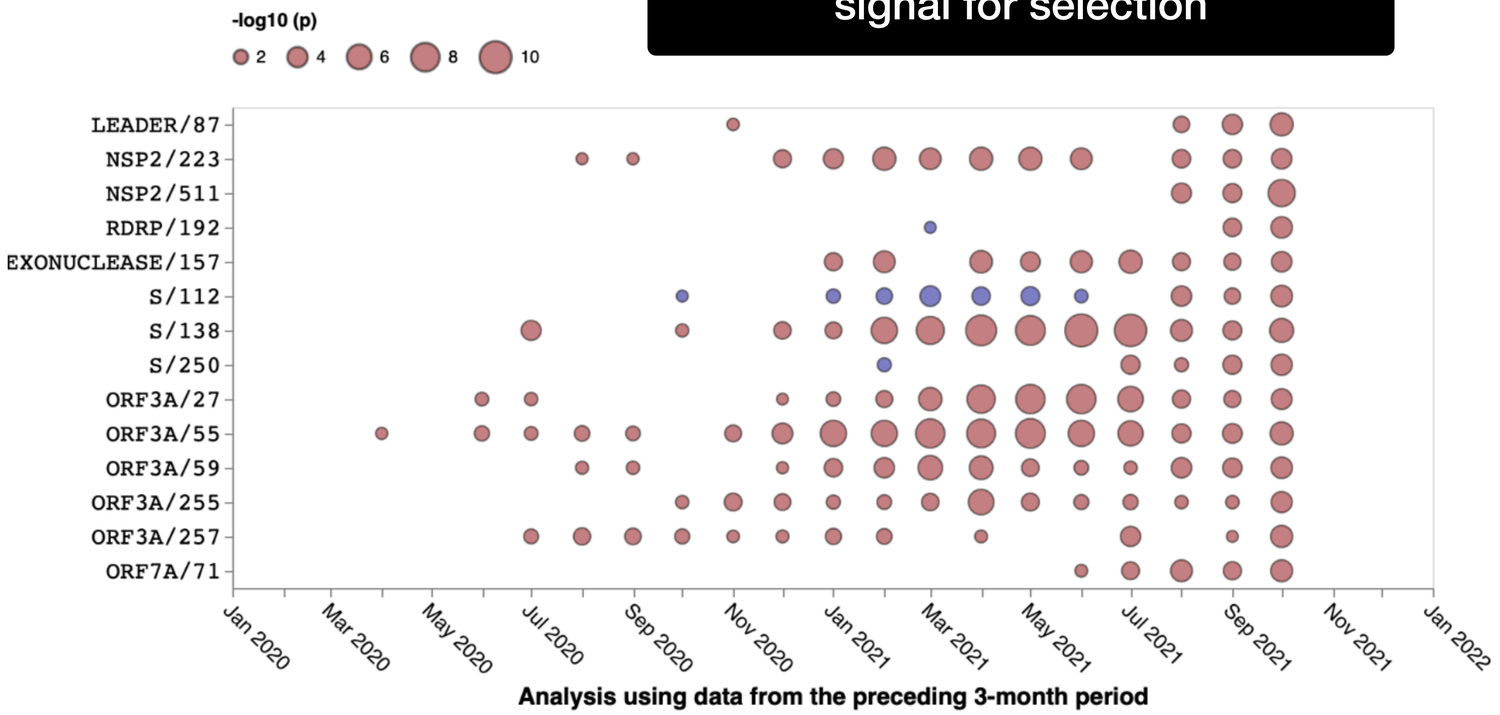
312 positively selected sites 20210801-20211031	**49↑169↓** change since 20210701-20210930	**3 [1-15]** time periods under selection (median, 95%)	**1390** positively selected at any point
380 negatively selected sites 20210801-20211031	**37↑155↓** change since 20210701-20210930	**3 [1-16]** time periods under selection (median, 95%)	**1503** negatively selected at any point

https://observablehq.com/@spond/selection-profile

December 2019 | November 2020 | May 2021

Continued evolution, complex selection dynamics, transition to endemic?

Sites with new or stronger recent signal for selection

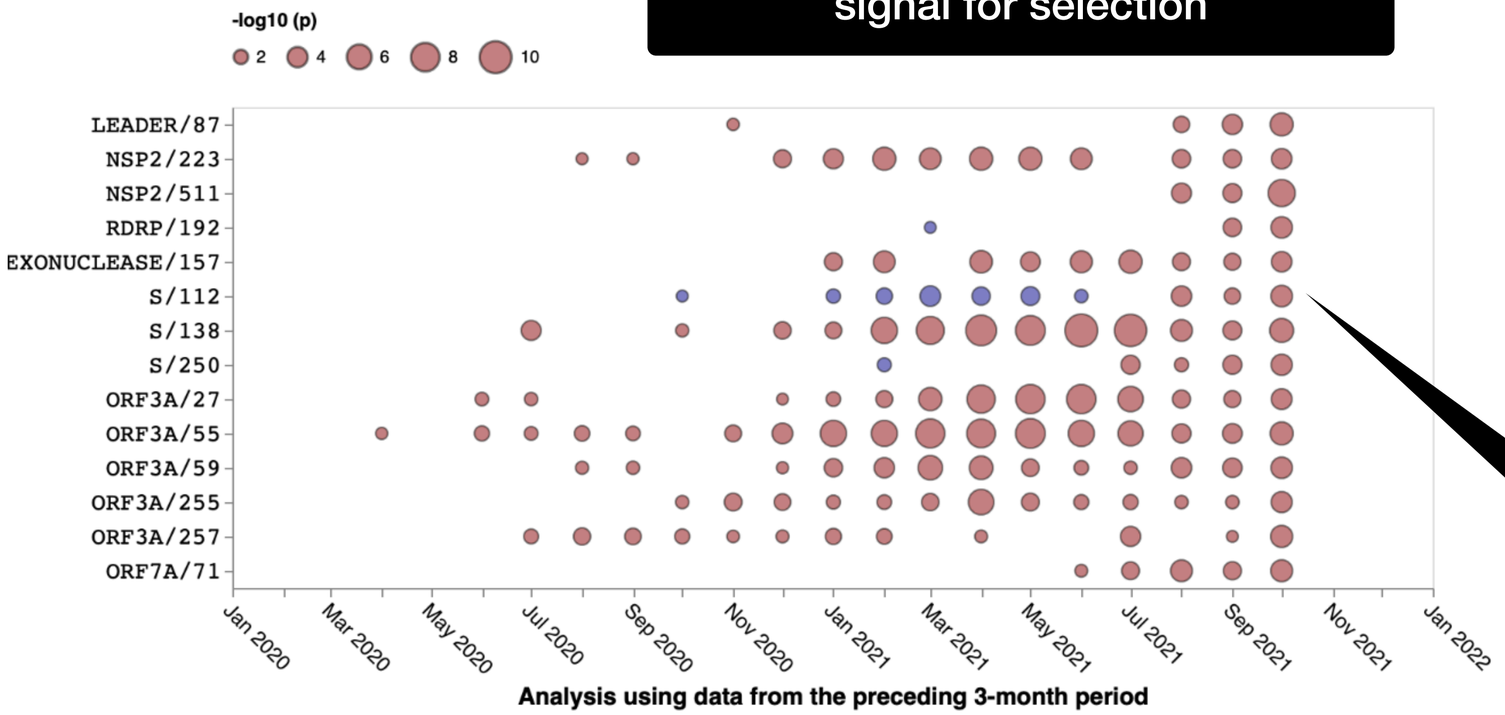


https://observablehq.com/@spond/selection-profile

December 2019 | November 2020 | May 2021

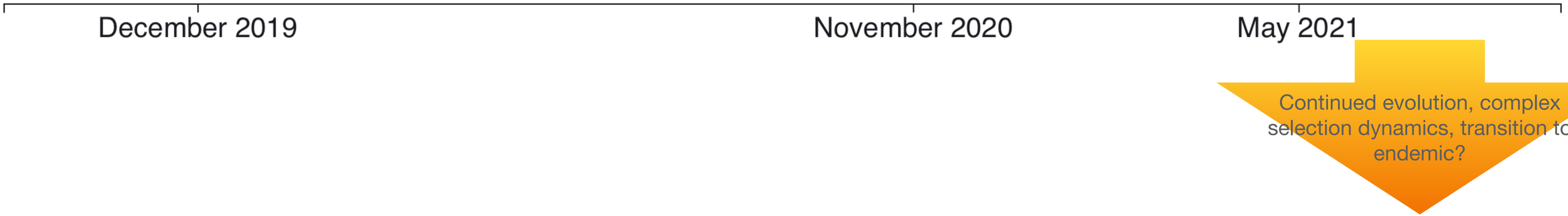
Continued evolution, complex selection dynamics, transition to endemic?

Sites with new or stronger recent signal for selection



S/112 used to be negatively selected, now is positively selected

https://observablehq.com/@spond/selection-profile



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

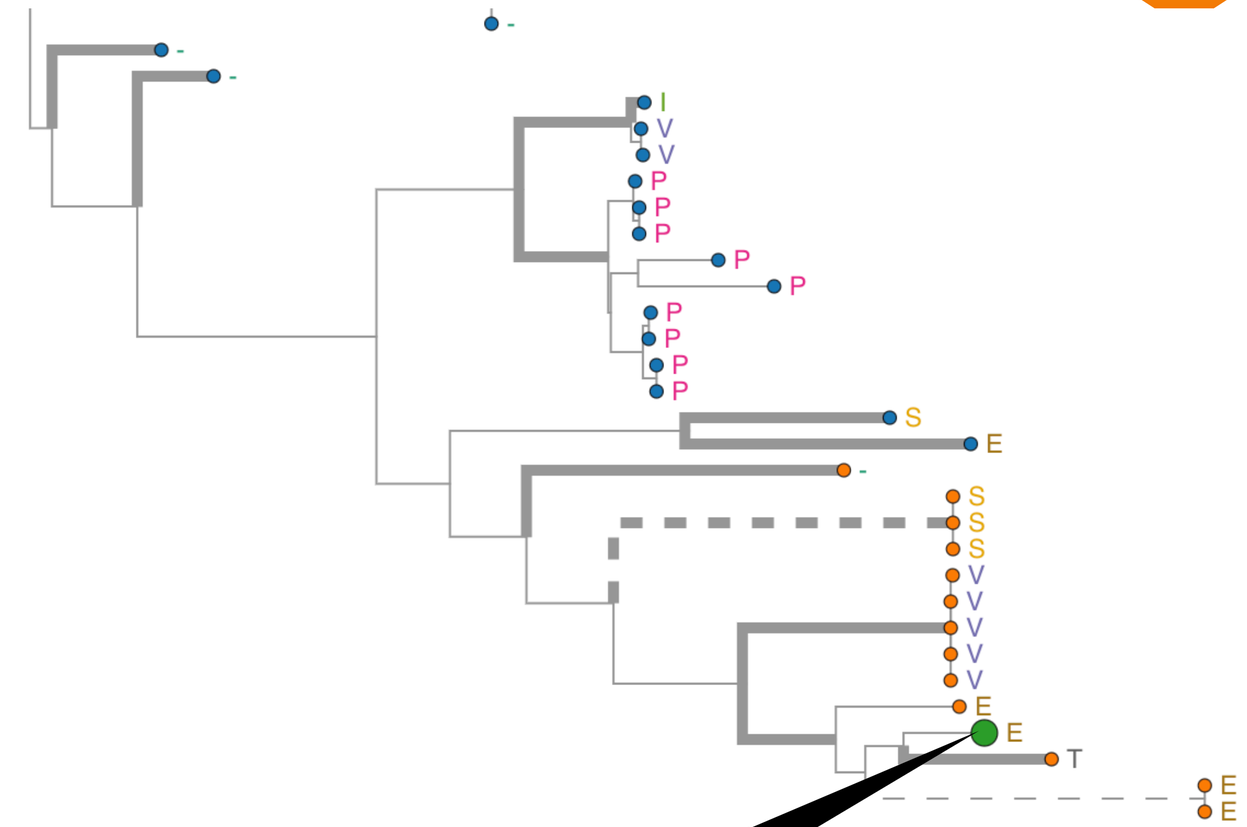
- Our evolutionary model uses inferred site-level biochemical property importance to impute evolutionary credibility.

https://observablehq.com/@spond/visualizing-selection-analysis-results-for-evolution-of-n

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.
- Our evolutionary model uses inferred site-level biochemical property importance to impute evolutionary credibility.



Predicting S/484 possible states based on nCOV evolution

https://observablehq.com/@spond/visualizing-selection-analysis-results-for-evolution-of-n

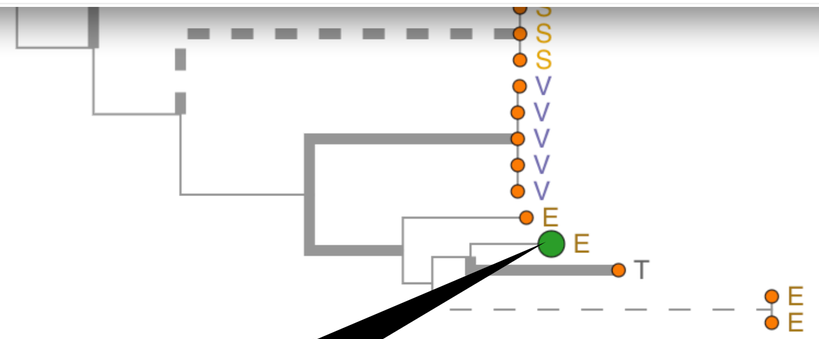
December 2019 | November 2020 | May 2021

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.
- Our evolutionary model uses inferred site-level biochemical property importanc
evolutionary credibi

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

Predicting S/484 possible states based on nCOV evolution

https://observablehq.com/@spond/visualizing-selection-analysis-results-for-evolution-of-n

December 2019 — November 2020 — May 2021

Continued evolution, complex selection dynamics, transition to endemic?

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 Spearkman 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)

https://observablehq.com/@spond/sars-cov-2-pvo

December 2019 | November 2020 | May 2021

Continued evolution, complex selection dynamics, transition to endemic?

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 Spearkman 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)

The set of “unusual” changes (compared to nCOV)

https://observablehq.com/@spond/sars-cov-2-pvo

December 2019

November 2020

May 2021

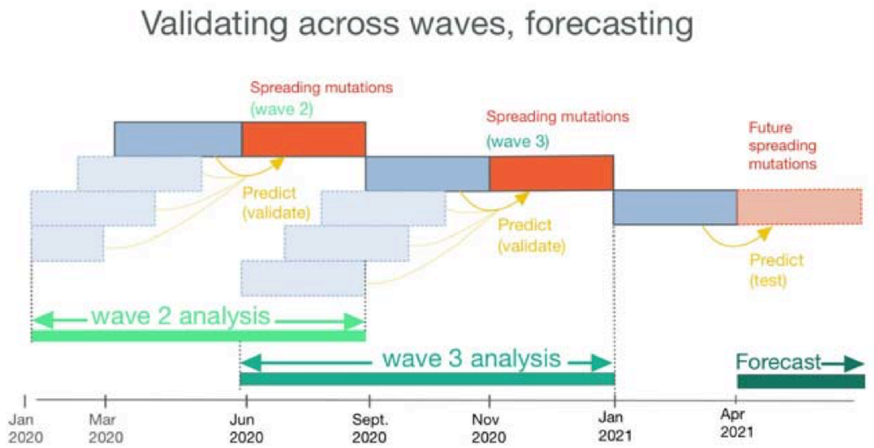
Continued evolution, complex selection dynamics, transition to endemic?

Maher et al “Predicting the mutational drivers of future SARS-CoV-2 variants of concern”, **MedRxiv**

December 2019 | November 2020 | May 2021

Continued evolution, complex selection dynamics, transition to endemic?

- Can we predict “short-term” evolutionary dynamics of the virus?

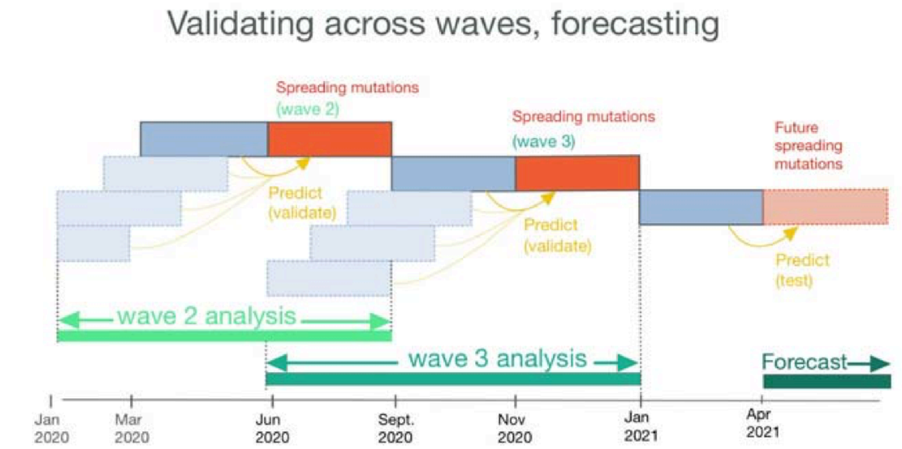


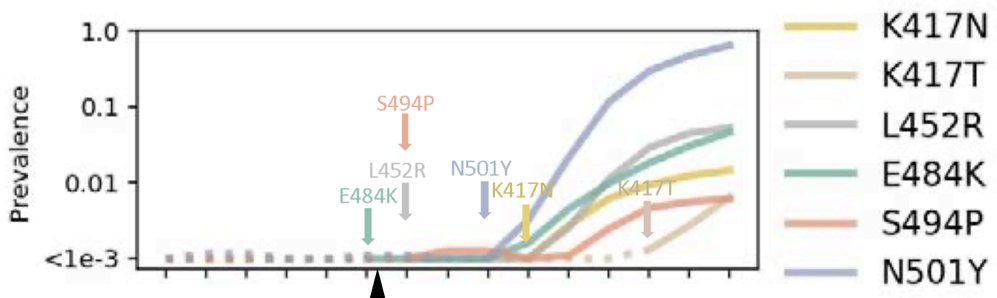
Maher et al “Predicting the mutational drivers of future SARS-CoV-2 variants of concern”, **MedRxiv**

December 2019 | November 2020 | May 2021

Continued evolution, complex selection dynamics, transition to endemic?

- Can we predict “short-term” evolutionary dynamics of the virus?
- We can **reliably** detect sites that will substantially increase in frequency in the next three months (AUC > 0.9)





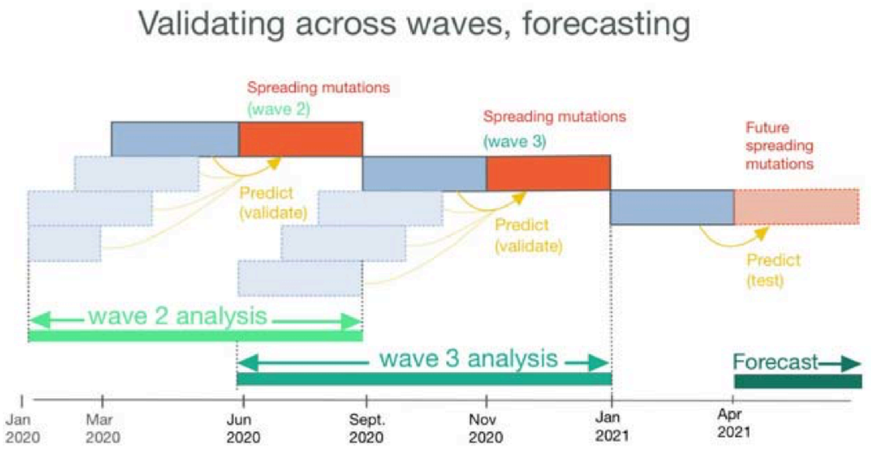
Variants detected before increasing in frequency

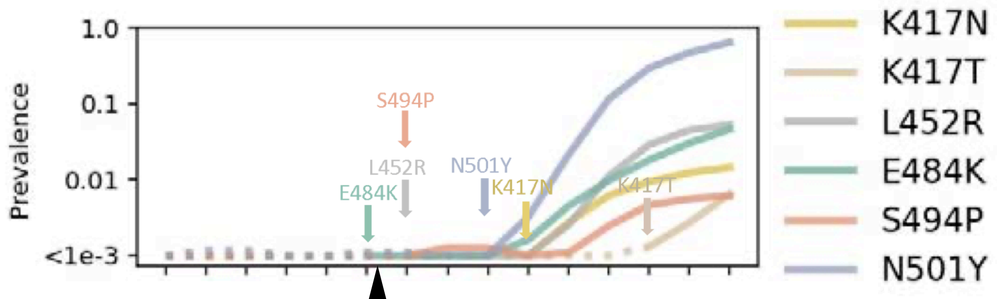
Maher et al “Predicting the mutational drivers of future SARS-CoV-2 variants of concern”, **MedRxiv**

December 2019 | November 2020 | May 2021

Continued evolution, complex selection dynamics, transition to endemic?

- Can we predict “short-term” evolutionary dynamics of the virus?
- We can **reliably** detect sites that will substantially increase in frequency in the next three months (AUC > 0.9)
- Most predictive metrics
  - Evidence of positive selection
  - Epidemiological data (**fraction of haplotypes with mutation**)





Variants detected before increasing in frequency

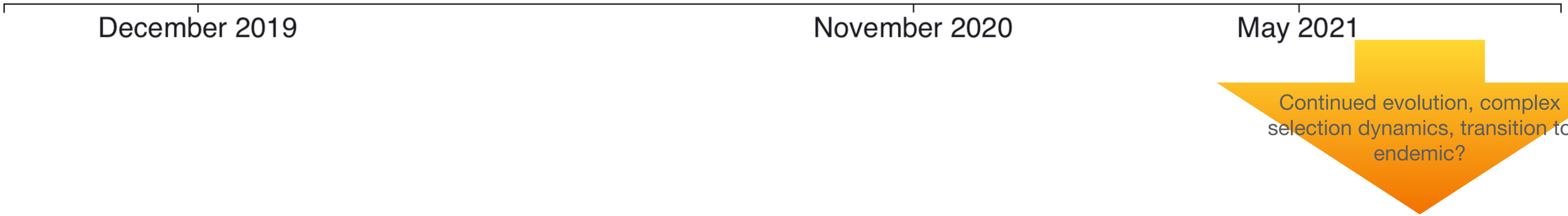
Maher et al “Predicting the mutational drivers of future SARS-CoV-2 variants of concern”, **MedRxiv**

December 2019

November 2020

May 2021

Continued evolution, complex selection dynamics, transition to endemic?



- Prioritizing a mutation as “adaptive” or “conserved” based on a simple combination of factors
  - Selective past and present of the site (positive, negative, neither)
  - Evolutionary credibility
  - Does this mutation appear in a large fraction of genomic contexts (haplotypes)?
  - …

December 2019 — November 2020 — May 2021

Continued evolution, complex selection dynamics, transition to endemic?

Coord	Gene/site	Recent	Type	Periods	Freq	Codon	AA	Predicted	HRank	AScore
28687	N/139	+	Alternating	14	0.0022	TTG→TTT	L→F	0	0.002	4.645
23398	S/613	+	Alternating	13	0.0039	CAG→CAT	Q→H	0	0	4.615
28687	N/139	+	Alternating	14	0.0018	TTG→TTC	L→F	0	0.006	4.613
17685	helicase/484	+	Alternating	8	0.0002	GTT→TTT	V→F	0	0.006	4.32
28678	N/136	+	Alternating	8	0.0008	GAG→GAT	E→D	0	0.006	4.319
26175	ORF3a/262	+	Alternating	7	0.0006	CCA→TCA	P→S	0	0.004	4.293
18969	exonuclease...	+	Alternating	8	0.0004	AAG→AAT	K→N	0	0.01	4.289
11740	nsp6/257	+	Alternating	8	0.0002	CAG→CAT	Q→H	0	0.011	4.277
25632	ORF3a/81	+	Alternating	8	0.0002	TGC→TTC	C→F	0	0.012	4.269
28678	N/136	+	Alternating	8	0.0001	GAG→CAG	E→Q	0	0.013	4.255
29299	N/343	+	Alternating	7	0.0003	GAT→CAT	D→H	0	0.008	4.253
26175	ORF3a/262	+	Alternating	7	0.0003	CCA→CTA	P→L	0	0.009	4.244
29347	N/359	+	Alternating	6	0.0008	GCA→TCA	A→S	0	0.005	4.24
29299	N/343	+	Alternating	7	0.0001	GAT→TAT	D→Y	0	0.01	4.235
24871	S/1104	+	Alternating	5	0.0041	GTA→TTA	V→L	0	0	4.235

https://observablehq.com/@spond/sc2-prioritization

- Better prediction of near-term evolutionary trajectories
- Prioritization of low frequency sites and site combinations for testing
- Detection of intra-host adaptation
- Analysis of selective forces associated with vaccine “breakthrough” infections

- Development of additional meta-signatures for “delta-like” and emergent lineages.
- Inter-operation with phenotypic annotation services and databases.
- Better prediction of phenotype from genotype.

Acknowledgements

Darren Martin
Steven Weaver
Oscar MacLean
Jordan Zehr
Alexander Lucaci
Stephen Shank
Hannah Kim
Tulio de Olivera
Amalio Telenti
David Robertson
Spyros Lytras

…many many others

NSF/NIH