Science Translational Medicine

RESEARCH ARTICLES

Cite as: M. C. Maher *et al.*, *Sci. Transl. Med.* 10.1126/scitranslmed.abk3445 (2022).

CORONAVIRUS

Predicting the mutational drivers of future SARS-CoV-2 variants of concern

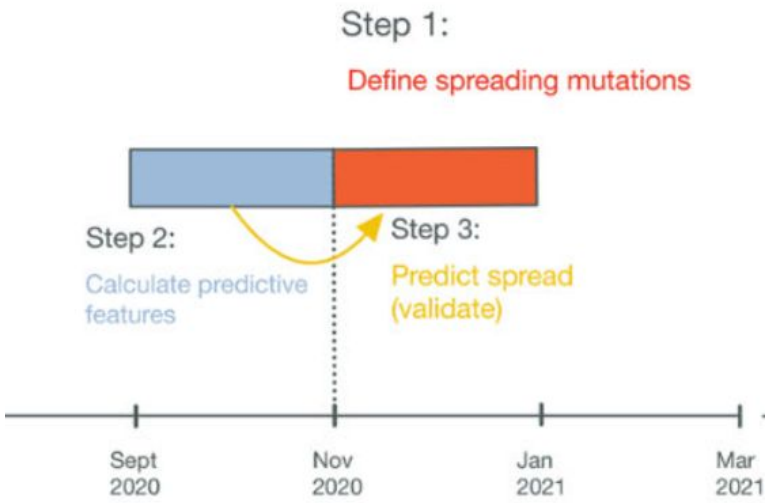
M. Cyrus Maher[1]*, Istvan Bartha[1], Steven Weaver[2], Julia di Iulio[1], Elena Ferri[1], Leah Soriaga[1], Florian A. Lempp[1], Brian L. Hie[3,4], Bryan Bryson[4,5], Bonnie Berger[6,7], David L. Robertson[8], Gyorgy Snell[1], Davide Corti[1], Herbert W. Virgin[1,9,10], Sergei L. Kosakovsky Pond[2], Amalio Telenti[1]*

[1]Vir Biotechnology, San Francisco, California 94158, USA [2]Department of Biology Institute for Genomics and Evolutionary Medicine Temple University, Philadelphia, PA 19122 [3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA. [5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [6]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [7]Computer Science and Artificial Intelligence Laboratory. Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow GS1 1QH, UK [9]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA [10]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX 75390, USA

Abstract What they set out to do

SARS-CoV-2 evolution threatens vaccine- and natural infection-derived immunity as well as the efficacy of therapeutic antibodies. To improve public health preparedness, we sought to predict which existing amino acid mutations in SARS-CoV-2 might contribute to future variants of concern. We tested the predictive value of features comprising epidemiology, evolution, immunology, and neural network-based protein sequence modeling, and identified primary biological drivers of SARS-CoV-2 intra-pandemic evolution. We found evidence that ACE2-mediated transmissibility and resistance to population-level host immunity has waxed and waned as a primary driver of SARS-CoV-2 evolution over time. We retroactively identified with high accuracy (area under the receiver operator characteristic curve, AUROC=0.92-0.97) mutations that will spread, at up to four months in advance, across different phases of the pandemic. The behavior of the model was consistent with a plausible causal structure wherein epidemiological covariates combine the effects of diverse and shifting drivers of viral fitness. We applied our model to forecast mutations that will spread in the future and characterize how these mutations affect the binding of therapeutic antibodies. These findings demonstrate that it is possible to forecast the driver mutations that could appear in emerging SARS-CoV-2 variants of concern. We validate this result against Omicron, showing elevated predictive scores for its component mutations prior to emergence, and rapid score increase across daily forecasts during emergence. This modeling approach may be applied to any rapidly evolving pathogens with sufficiently dense genomic surveillance data, such as influenza, and unknown future pandemic viruses.

Main goal Predicting spreading mutations



Step 1: What are spreading mutations?

- Comparing the ratio of sequences containing the mutation in **three-months windows before and after the date of interest** for **each country separately**

  → p-values → adjustment for multiple testing (q)

  → **mutations with q < 0.05 for any country kept**

- Additional empirical criteria
  - A fold-change (FC) of at least 10 in at least one country
  - A FC of at least 2 in at least 3 countries
  - Minimum global ratio of at least 0.1% in the later time window

Main goal Predicting spreading mutations

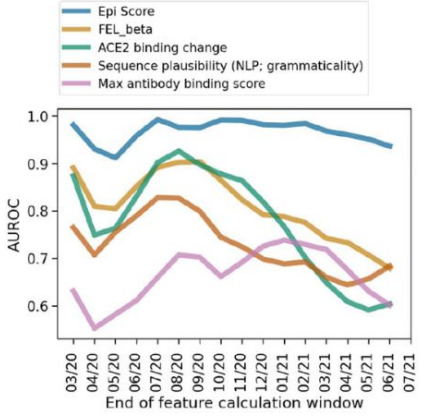
Table 1. Summary of analytical features. A total of 48 parameters for 14 variables were created for 5 feature groups. These features capture evolutionary, immune, epidemiologic, transmissibility, and language model predictors of mutation spread. A detailed description of all parameters is included in **data file S1**.

Feature group	Variable	Meaning	Source or reference	Number of parameters
Evolution	Positive selection (FEL, MEME)	Parameters from Fixed Effects Likelihood (FEL) and Mixed Effects Model of Evolution (MEME)	HyPhy (*19*)	11
	Codon-SHAPE	RNA SHAPE constraint	Manfredonia *et al.* 2020 (*32*)	3
	Viral entropy	Shannon entropy at each codon position for an amino acid site	This work	3
Immune	CD8 epitope escape	The frequency of SARS CoV-2 mutations in cytotoxic lymphocyte (CTL) epitopes	Agerer *et al.* 2021 (*15*)	1
	CD8 response	The percent and average CD8+ T cell response to an epitope in patients	Tarke *et al.* 2021 (*33*)	2
	CD4 response	The percent and average CD4+ T cell response to an epitope in patients	Tarke *et al.* 2021 (*33*)	2
	Antibody binding score	The estimated percent contribution of a site to binding of the indicated antibody, as estimated by Molecular Operating Environment (MOE)	This work	17
	Maximum escape fraction in vitro	The maximum escape fraction across all conditions for that mutation	Greaney *et al.* 2021 (*34*)	1
Epidemiology	Variant frequency	The percent of sequences with the mutation	Calculated from GISAID (*2*)	1
	Fraction of unique haplotypes	The fraction of unique Spike haplotypes in which a mutation is observed	Calculated from GISAID (*2*)	1
	Number of countries	The number of countries where it has been observed.	Calculated from GISAID (*2*)	1
	Epi Score	The exponentially weighted mean rank across the other epidemiology variables	Calculated from GISAID (*2*)	1
Transmissibility	RBD expression change	Change in RBD expression due to the mutation	Starr *et al.* 2020 (*13*)	1
	ACE2 binding change	The change in binding affinity for ACE2	Starr *et al.* 2020 (*13*)	1
Language model	Language model	Grammaticality and semantic change of a mutation	Hie *et al.* 2021 (*17*)	2

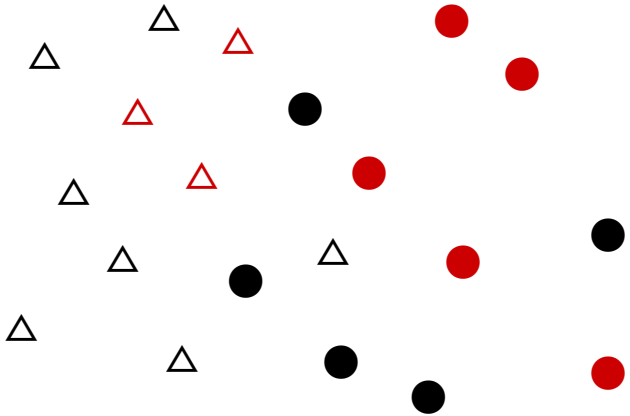
Steps 2 & 3:

What are predictive features?
How well do they predict spreading?

Metric	RBD AUROC	Spike AUROC	Feature group
Epi Score	0.989	0.964	Epidemiology
FEL	0.855	0.837	Evolution
ACE2 binding change	0.848	0.553	Transmissibility
Sequence plausibility (grammaticality)	0.817	0.756	Language model
Maximal antibody binding score	0.709	0.575	Immune



Main goal Predicting spreading mutations



△: non-spreading predicted as non-spreading
△: non-spreading predicted as spreading
●: spreading predicted as non-spreading
●: spreading predicted as spreading

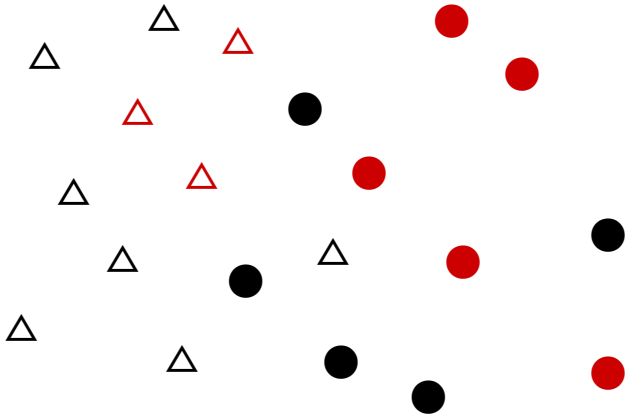
Steps 2 & 3:

What are predictive features? How well do they predict spreading?

Metric	RBD AUROC	Spike AUROC	Feature group
Epi Score	0.989	0.964	Epidemiology
FEL	0.855	0.837	Evolution
ACE2 binding change	0.848	0.553	Transmissibility
Sequence plausibility (grammaticality)	0.817	0.756	Language model
Maximal antibody binding score	0.709	0.575	Immune

- **univariate "model":** use **thresholds** to differentiate between non-spreading/spreading mutations

Main goal Predicting spreading mutations



Sensitivity: ●/(●+●) = 0.5

△: non-spreading predicted as non-spreading
△: non-spreading predicted as spreading
●: spreading predicted as non-spreading
●: spreading predicted as spreading

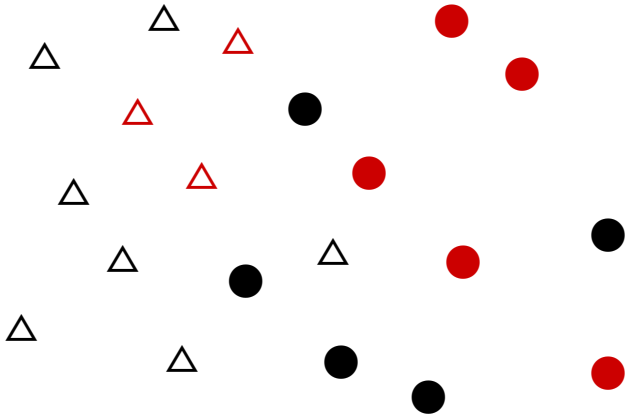
Steps 2 & 3:

What are predictive features? How well do they predict spreading?

Metric	RBD AUROC	Spike AUROC	Feature group
Epi Score	0.989	0.964	Epidemiology
FEL	0.855	0.837	Evolution
ACE2 binding change	0.848	0.553	Transmissibility
Sequence plausibility (grammaticality)	0.817	0.756	Language model
Maximal antibody binding score	0.709	0.575	Immune

- **univariate "model":** use **thresholds** to differentiate between non-spreading/spreading mutations
- **sensitivity:** fraction of spreading mutations correctly forecast (high with lenient thresholds)

Main goal Predicting spreading mutations



Sensitivity: ●/(●+●) = 0.5
Positive predictive value: ●/(●+△) = 0.625

△: non-spreading predicted as non-spreading
△: non-spreading predicted as spreading
●: spreading predicted as non-spreading
●: spreading predicted as spreading

Steps 2 & 3:

What are predictive features? How well do they predict spreading?

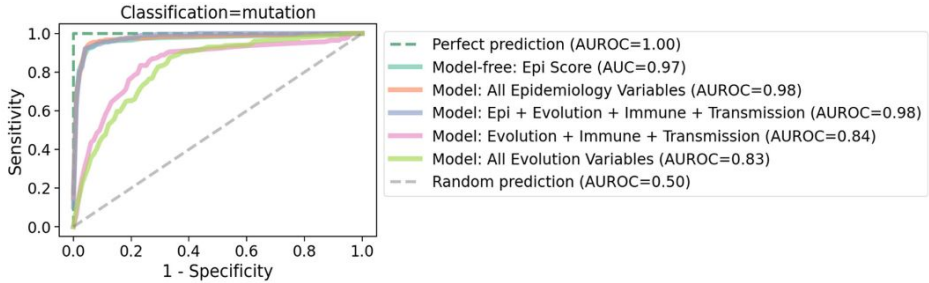
Metric	RBD AUROC	Spike AUROC	Feature group
Epi Score	0.989	0.964	Epidemiology
FEL	0.855	0.837	Evolution
ACE2 binding change	0.848	0.553	Transmissibility
Sequence plausibility (grammaticality)	0.817	0.756	Language model
Maximal antibody binding score	0.709	0.575	Immune

- **univariate "model":** use **thresholds** to differentiate between non-spreading/spreading mutations
- **sensitivity:** fraction of spreading mutations correctly forecast (high with lenient thresholds)
- **positive predictive value:** fraction of predicted spreading mutations that are correct (high with strict thresholds)

Main goal Predicting spreading mutations

Steps 2 & 3: Should features be combined?

- **univariate "model":** use **thresholds** to differentiate between non-spreading/spreading mutations
- **multivariate model:** logistic regression



Final model:

- use Epi Score to sort mutations
- predict the top 5% of mutations as spreading

Final model What it does and does not do?

- use **Epi Score** to sort mutations
- predict the top 5% of mutations as spreading

Variant frequency	The percent of sequences with the mutation	Calculated from GISAID (*2*)
Fraction of unique haplotypes	The fraction of unique Spike haplotypes in which a mutation is observed	Calculated from GISAID (*2*)
Number of countries	The number of countries where it has been observed.	Calculated from GISAID (*2*)
Epi Score	The exponentially weighted mean rank across the other epidemiology variables	Calculated from GISAID (*2*)

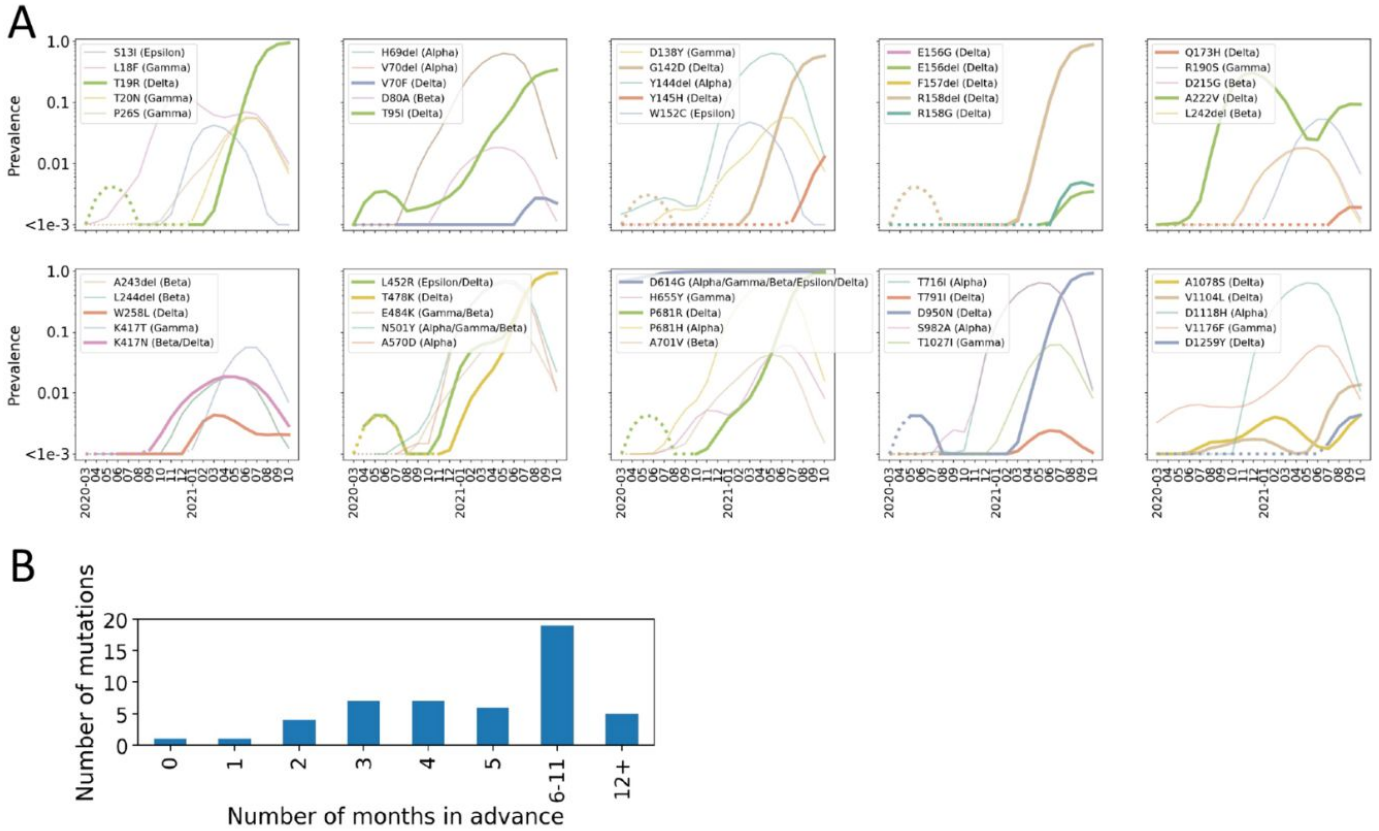
Those mutations will be categorized as spreading that

- are already present in many sequences in GISAID,
- are already present in many unique Spike haplotypes in GISAID sequences,
- are already present in many countries.

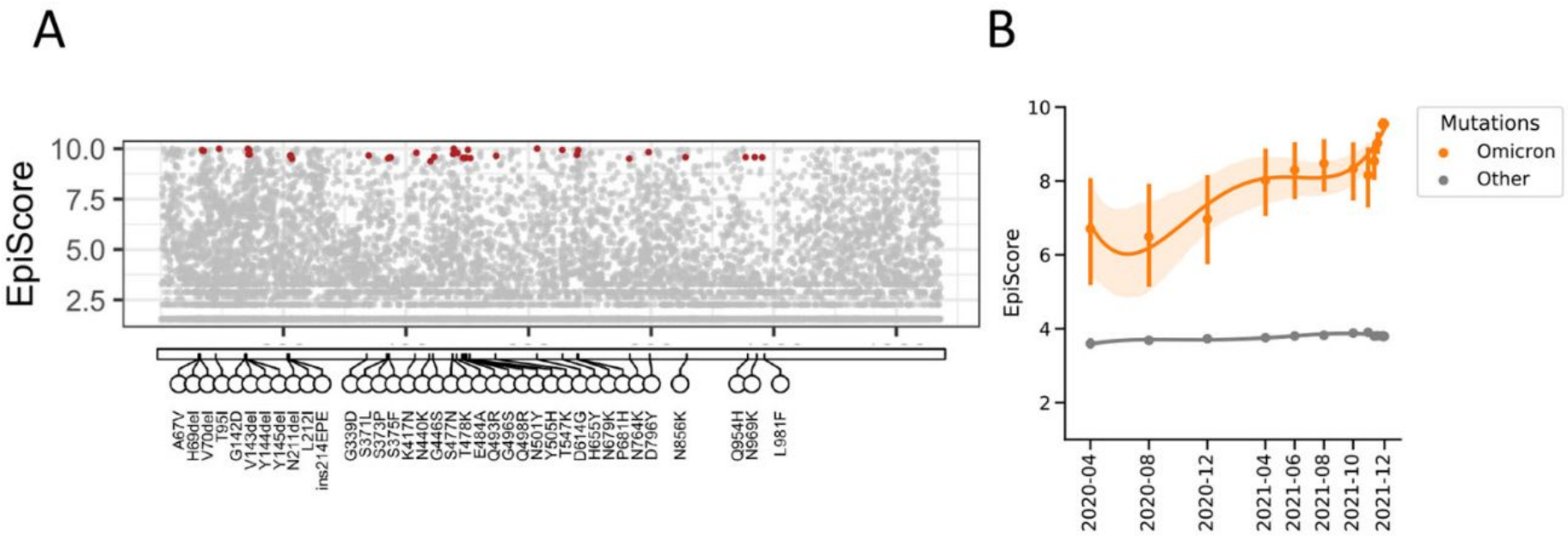
The model basically **detects spreading sooner** than we do manually.

The model cannot forecast the spreading of mutations that are not yet present in a substantial number of sequences based on biological considerations. It also cannot forecast the simultaneous spreading of multiple mutations (variants).

Validation Retrospective forecasting for VoC

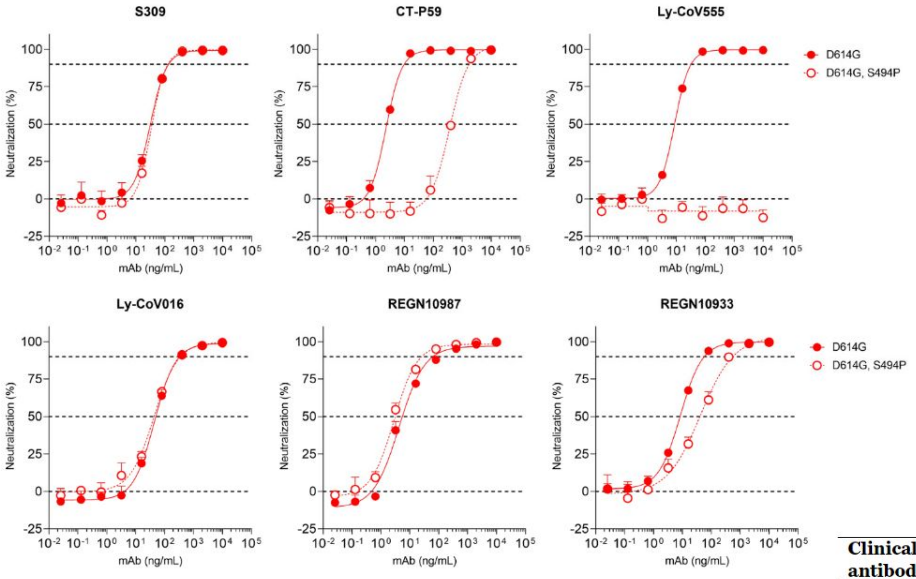


Validation Omicron mutations



Value of predictions Prioritize mutations for functional screening

A



B

Antibody	EC50 (ng/mL)		Fold-change
	D614G	D614G, S494P	
S309	35.6	40.7	1.1
CT-P59	2.3	393.0	172.1
Ly-CoV555	8.2	>10,000	>1213
Ly-CoV016	47.1	42.6	0.9
REGN10987	6.5	3.8	0.6
REGN10933	6.7	38.6	5.9

Clinical therapeutic antibody	Forecasted mutations in epitopes
VIR-7831 (sotrovimab)	A344S†, R346K†
LY-CoV016 (etesevimab)	K417T‡, K417N*, L455F‡
REGN10987 (imdevimab)	R346K†, K444N*, G446V*
LY-CoV555 (bamlanivimab)	L452R*, L452Q‡, V483F†, E484K*, E484Q*, F490S*, S494L‡, S494P*
REGN10933(casirivimab)	K417T*, K417N*, L455F*, G476S*, S477I‡, T478K‡, E484K*, E484Q*, F490S*
CT-P59	K417T‡, K417N†, L452R*, L452Q‡, L455F‡, E484K*, E484Q‡, F490S‡, S494L‡, S494P‡