D888–D897 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 11 October 2021*
https://doi.org/10.1093/nar/gkab921

VarEPS: an evaluation and prewarning system of known and virtual variations of SARS-CoV-2 genomes

Qinglan Sun[1,3,†], Chang Shu[1,2,†], Wenyu Shi[1,3], Yingfeng Luo[1,2,7], Guomei Fan[1,3], Jingyi Nie[1,2,7], Yuhai Bi[4], Qihui Wang[4], Jianxun Qi[4], Jian Lu[5], Yuanchun Zhou[6], Zhihong Shen[6], Zhen Meng[6], Xinjiao Zhang[1,3], Zhengfei Yu[1,3], Shenghan Gao[1,2,*], Linhuan Wu[1,3,*], Juncai Ma[1,2,3,*] and Songnian Hu[1,2,7,*]

[1]Microbial Resource and Big Data Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China, [2]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China, [3]Chinese National Microbiology Data Center (NMDC), Beijing 100101, China, [4]CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China, [5]State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, School of Life Sciences, Peking University, Beijing 100871, China, [6]Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China and [7]University of Chinese Academy of Sciences, Beijing 100049, China

Received August 13, 2021; Revised September 18, 2021; Editorial Decision September 26, 2021; Accepted September 28, 2021

ABSTRACT

The genomic variations of SARS-CoV-2 continue to emerge and spread worldwide. Some mutant strains show increased transmissibility and virulence, which may cause reduced protection provided by vaccines. Thus, it is necessary to continuously monitor and analyze the genomic variations of SARS-COV-2 genomes. We established an evaluation and prewarning system, SARS-CoV-2 variations evaluation and prewarning system (VarEPS), including known and virtual mutations of SARS-CoV-2 genomes to achieve rapid evaluation of the risks posed by mutant strains. From the perspective of genomics and structural biology, the database comprehensively analyzes the effects of known variations and virtual variations on physicochemical properties, translation efficiency, secondary structure, and binding capacity of ACE2 and neutralizing antibodies. An AI-based algorithm was used to verify the effectiveness of these genomics and structural biology characteristic quantities for risk prediction. This classifier could be further used to group viral strains by their transmissibility and affinity to neutralizing antibodies. This unique resource makes it possible to quickly evaluate the variation risks of key sites, and guide the research and development of vaccines and drugs. The database is freely accessible at www.nmdc.cn/ncovn.

INTRODUCTION

As an RNA virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has a relatively high mutation rate (1) with a mean annual average evolutionary rate of 1×10^{-3} substitutions per base per year under conditions of neutral genetic drift (2). Since the initial outbreak in December 2019, a substantial number of SARS-CoV-2 variants have emerged. As of August 2021, a total of 2 635 714 SARS-CoV-2 genome sequences have become available in the Global Initiative of Sharing All Influenza Data (GISAID database), and 29 212 mutations have accumulated over the past year and a half. However, most mutations in SARS-CoV-2 occur at a very low frequency and cause no significant effect on the virus (3). Only a small number of mutations, especially those in the spike (S) protein, can change the infectivity of the virus and hence increase transmission or reduce the binding affinity of the S protein receptor-binding domain (S-RBD) for neutralizing antibodies. For instance, a point mutation in the S protein, D614G, shifts the conformation of the S protein toward an angiotensin-converting enzyme 2 (ACE2)-binding fusion-competent state and hence enhances SARS-CoV-2 infectivity in human lung cells (4). Research using computational simulation has suggested that some

[*]To whom correspondence should be addressed. Tel: +86 10 64807385; Email: wulh@im.ac.cn
Correspondence may also be addressed to Shenghan Gao. Email: gaoshh@im.ac.cn
Correspondence may also be addressed to Juncai Ma. Email: ma@im.ac.cn
Correspondence may also be addressed to Songnian Hu. Email: husn@im.ac.cn
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

mutations, including Q24T, T27D/K/W, D30E, H34S7T/K, E35D, Q42K, L79I/W, R357K and R393K in ACE2, and L455D/W, F456K/W, Q493K, N501T and Y505W in S-RBD, increase the binding affinity between ACE2 and S-RBD. Experimental evidence has shown that these *in silico* simulations are highly accurate (5).

Efficient and accurate diagnosis of COVID-19 is crucial for controlling the pandemic in early time. Reverse transcription polymerase chain reaction (RT-PCR) technology is the most widely used among the common diagnostic methods (6). Mutations at the probe or primer sites could have effects on the accuracy of diagnosis, such as loss of primer efficacy. As a result, continued surveillance of genomic mutation is crucial for disease control and vaccine and drug studies.

We here established the variations evaluation and prewarning system (VarEPS) and conducted a comprehensive analysis of the effects of variants on physicochemical properties, translation efficiency, secondary structure, difficulty in developing variations, binding capacity of ACE2 and binding capacity of neutralizing antibodies. To our knowledge, this is the most comprehensive analysis and risk evaluation of SARS-CoV-2 genome variants. Instead of the classical risk evaluation by variation frequency, we followed a new perspective on the effect of mutations on protein structure and function. Moreover, we constructed two random forest classifiers to verify the effectiveness of these characteristic quantities for accurate risk evaluation. This AI-based classifier can be used to accurately group strains by their transmissibility and affinity to neutralizing antibodies.

More importantly, we analyzed not only known variants but also virtual variants; as a result, by closely observing newly submitted genome sequences, we can identify emerging dangerous variants at an early stage. This platform can also yield vital information for virologists using pseudoviruses to test vaccines and drugs. Currently, VarEPS is the only database which provides these unique resources on virtual variants and is thus expected to be of great interest for virologists, especially those involved in vaccine and drug development.

DATABASE INTERFACE AND FEATURES

Database interface

The web interface of VarEPS is composed of five main sections: ‘Virus and variation’, ‘Binding ability evaluation’, ‘Primer efficacy evaluation’, ‘Statistics’ and ‘Analysis tools’ (Figure 1). The ‘Virus and variation’ section starts with search interfaces for metadata attributes of viral sequences and nucleotide variants. The resulting viral sequences with associated metadata are displayed as a table, including lineage, single nucleotide polymorphisms (SNP) number, and variation information for both nucleotides and amino acids. Each viral sequence is linked to an individual page containing all of the related mutations and primer evaluation results. A machine learning model is used to give an overall risk level prediction for each virus. The query on nucleotide variation returns a variant list with metadata related to the number of variations and the associated amino acid mutations. Each variation is linked to a page containing graphs of distribution over time and by country, and listing related viral sequences.

The ‘Binding ability evaluation’ section assesses the risk level of each virus variant. Variants may be queried and browsed by their location on genes, lineages and antibody binding sites. After query by different metadata, a list containing all amino acid mutations is returned. Antibody affinity, binding stability with ACE2, risk of amino acid substitution, and the first-seen and last-seen time are calculated and displayed. Each amino acid variation is linked to a page containing details of these values or the risk level.

The ‘Primer efficacy evaluation’ section assesses how mutations affect primer design for RT-PCR. Primer information is obtained from the USA Centers for Disease Control and Prevention (CDC), the Chinese Center for Disease Control and Prevention (CDC China), the World Health Organization (WHO) and others. If mutations are present in the 5′- and 3′- end, the primers might be of low specificity or lose efficacy entirely.

Online data analysis pipelines

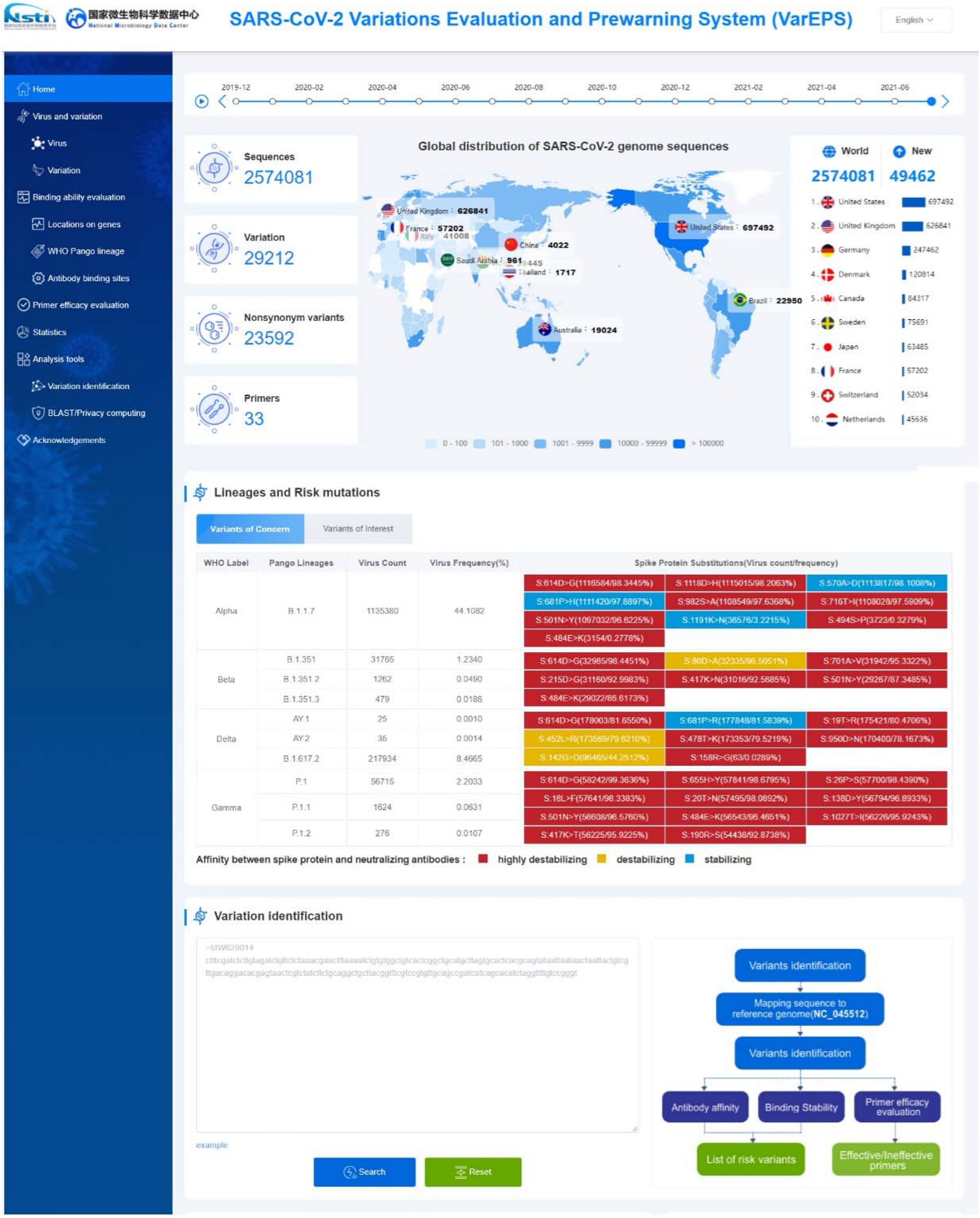
Online analysis tools are provided for users to submit sequences for variation analysis. Sequences are aligned against the reference genome (NC_045512.2) using NUCmer from the MUMmer package (7). Thereafter, a catalog of all SNPs and indels internal to the reference genome is generated. The system evaluates variants and generates risk level results by assessing amino acid substitution, binding affinity for ACE2 and secondary structure change. For variants of the S-RBD, the affinity with 15 neutralizing antibodies under development is calculated. Nucleotide mismatches with primers or probes are reported to warn of possible false negative results in diagnostic detection of SARS-CoV-2 by real-time RT-PCR. An evaluation report of the submitted virus is sent to users via e-mail after all analyses are complete.

Statistics

A statistics page organized by ‘Lineage’, ‘Variations’ and ‘Primer’ provides an overview of statistical analysis of variants. The ‘Lineage’ page displays the distribution of different lineages by country and through time. The ‘Variations’ page gives a set of graphs on variant distribution and risk level of different lineages. The ‘Primer’ page lists primer evaluation results of different lineages. Interactive interfaces are provided to allow the user to further explore the features of various groups.

DATA CONTENT AND ANALYSIS

We calculate the occurrence of each mutation site in nucleotide/amino acid variants against the reference sequence. Currently, there are 29 212 variants observed on nearly 30 000 nucleotides of the SARS-CoV-2 whole length genome sequence. However, many variants are of a very low frequency. Among the 29 212 nucleotide variants, 4672 (16.0%) sites occur <10 times and 10 920 (37.4%) sites occur <50 times. Only 1650 (5.7%) sites occur >2600 times (with a frequency of 0.1%) and 33 (0.1%) sites occur >24 000 times (with a frequency of 1%) (Figure 2A). The SARS-CoV-2 mutation rate is vital to determining how quickly the transmissibility of a virus changes and immune evasion occurs.



Sequences 2574081

Variation 29212

Nonsynonym variants 23592

Primers 33

World 2574081

New 49462

1.	United States	697492
2.	United Kingdom	626841
3.	Germany	247462
4.	Denmark	120814
5.	Canada	84317
6.	Sweden	75691
7.	Japan	63485
8.	France	57202
9.	Switzerland	52034
10.	Netherlands	45636

Lineages and Risk mutations

WHO Label	Pango Lineages	Virus Count	Virus Frequency(%)	Spike Protein Substitutions(Virus count/frequency)		
Alpha	B.1.1.7	1135380	44.1082	S:614D>G(1116584/98.3445%)	S:1118D>H(1115015/98.2063%)	S:570A>D(1113817/98.1008%)
				S:681P>H(1111420/97.8897%)	S:982S>A(1108549/97.6368%)	S:716T>I(1108028/97.5909%)
				S:501N>Y(1097032/96.6225%)	S:1191K>N(36576/3.2215%)	S:494S>P(3723/0.3279%)
				S:484E>K(3154/0.2778%)		
Beta	B.1.351	31765	1.2340	S:614D>G(32985/98.4451%)	S:80D>A(32335/96.5051%)	S:701A>V(31942/95.3322%)
	B.1.351.2	1262	0.0490	S:215D>G(31160/92.9983%)	S:417K>N(31016/92.5685%)	S:501N>Y(29267/87.3485%)
	B.1.351.3	479	0.0186	S:484E>K(29022/86.6173%)		
Delta	AY.1	25	0.0010	S:614D>G(178003/81.6550%)	S:681P>R(177848/81.5839%)	S:19T>R(175421/80.4706%)
	AY.2	35	0.0014	S:452L>R(173569/79.6210%)	S:478T>K(173353/79.5219%)	S:950D>N(170400/78.1673%)
	B.1.617.2	217934	8.4665	S:142G>D(96465/44.2512%)	S:158R>G(63/0.0289%)	
Gamma	P.1	56715	2.2033	S:614D>G(58242/99.3636%)	S:655H>Y(57841/98.6795%)	S:26P>S(57700/98.4390%)
	P.1.1	1624	0.0631	S:18L>F(57641/98.3383%)	S:20T>N(57495/98.0892%)	S:138D>Y(56794/96.8933%)
				S:501N>Y(56608/96.5760%)	S:484E>K(56543/96.4651%)	S:1027T>I(56226/95.9243%)
	P.1.2	276	0.0107	S:417K>T(56225/95.9225%)	S:190R>S(54438/92.8738%)	

Affinity between spike protein and neutralizing antibodies : 🟥 highly destabilizing 🟨 destabilizing 🟦 stabilizing

Variation identification

```
>MW629014
ctttcgatctcttgtagatctgttctctaaacgaactttaaaatctgtgtggctgtcactcggctgcatgcttagtgcactcacgcagtataattaataactaattactgtcg
ttgacaggacacgagtaactcgtctatcttctgcaggctgcttacggtttcgtccgtgttgcagccgatcatcagcacatctaggtttgtccgggt
```

Figure 1. Features of the variations evaluation and prewarning system (VarEPS) portal. We show a global distribution of genome sequences by time frame and geography. The risk level and frequency of characteristic variants of each lineage are listed. Users can submit a sequence for variation analysis directly on the homepage.

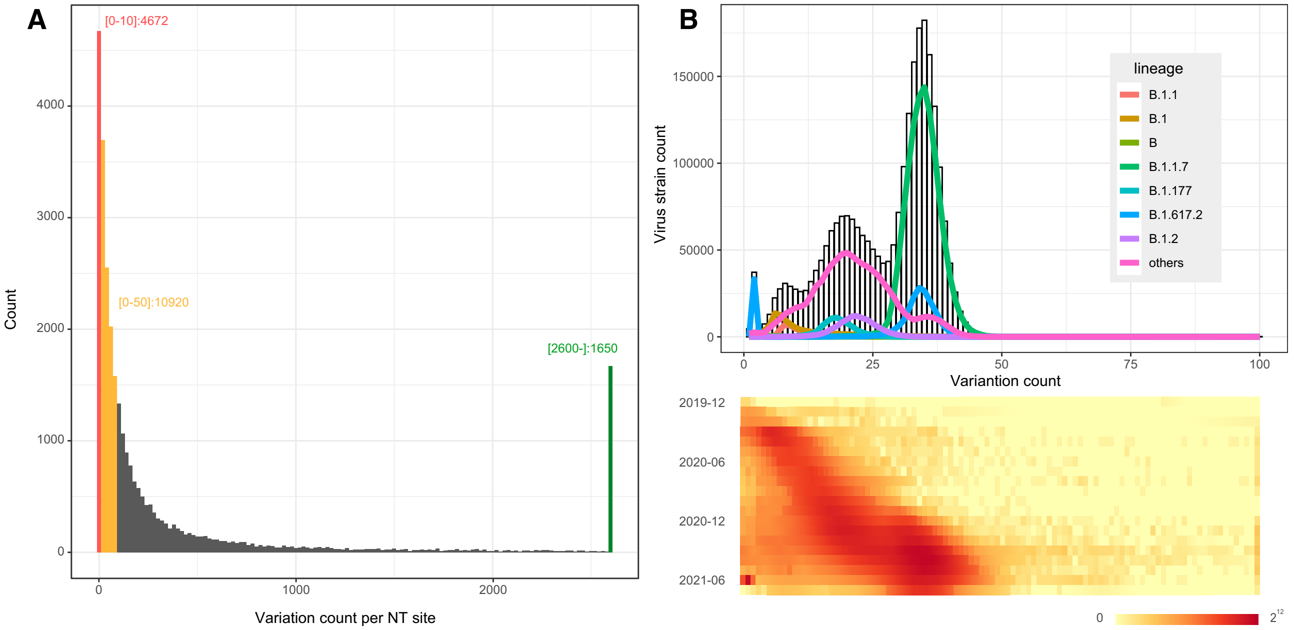


Figure 2. Statistics of nucleotide mutation numbers in SARS-CoV-2 genomes. (**A**) Histogram of the mutation count at all nucleotide positions. Red, orange and green bars refer to the frequency of mutation count below 10, below 50 and above 2600, respectively. (**B**) Histogram of total mutation count in one strain. The heatmap shows the distribution of total mutation count in each month. Mutation counts are accumulating over time and coordinate with lineages.

The mean annual mutation rate is reported to be 1 × 10^{-3} substitutions per base per year (2), and apparently, the observed mean mutation occurrence rate is consistent with the estimated rates and is closely associated with lineage (Figure 2B).

Among amino acid variants, the most frequent variant is D614G. The next most frequent, N501Y, is located in the S-RBD, whereas the frequency of all other S-RBD variants is <10% (Table 1). Still, a large number of high frequency mutations are located outside of the S protein. Variants that appear in viral populations with a high frequency or that are located in domains with critical effects on viral structure or function should be given our utmost attention.

SARS-CoV-2 S-RBD is the molecular target for most SARS-CoV-2 vaccines and antibodies currently in use or under development. We compared key amino acid mutations (the top 20 most frequent variants) in the S-RBD for their effects on S protein affinity with neutralizing antibodies and ACE2 (Figure 3). The simulated results showed that the most frequent variants reduced the binding affinity of the S protein for neutralizing antibodies. This result should be followed up with *in vivo* experiments to test the simulation results and examine the effects. Other variants (e.g. L452R and K417T) exhibited increased affinity with ACE2, indicating enhanced infectivity of these variants. Combined with the distributions with time span, it is critical to pay close attention to the risk presented by emerging variants that rapidly increase in frequency.

Apart from the existing mutations, this platform allows evaluation of new mutations as they appear in the future. Evaluating the risk level of virtual mutations could facilitate drug and/or vaccine development. From the simulation results (Figure 4), we estimated that antibody affinity will be reduced as a result of most of these virtual mutations. Binding stability to ACE2 will also be affected by mutations in some key positions (e.g. 345, 413, 520 and 522).

L452R is an important mutation that has commonly appeared in the recently prevalent Alpha, Delta, Epsilon, Iota and Kappa strains, and it is reported that the variants can reduce sensitivity to neutralizing antibodies (8). This mutation may increase affinity for ACE2 receptors and accordingly increase infectivity (9). Consistent with these experimental results, our prewarning system results indicated that the variation may be associated with increased infectivity and decreased affinity with some neutralizing antibodies. Additionally, we predicted all possible variants at this site; the data revealed that the risk level for some variants was even higher than the currently widespread L452R, including L452Q, which is one of the characteristic variants of Lambda strains. Others were virtual variants that have not yet, such as L452A, L452N and L452D. Emerging variants should be closely monitored for such mutations with high predicted risk levels.

Finally, we list variants that could affect the performance of the primers recommended by WHO, CDC and CDC China. These data are organized by number of mismatched nucleotides for different lineages (Figure 5). Most of these mismatches occur at the first nucleotide of the 3′ end for Alpha strains. However, the number of affected viruses is very low. Considering the high percentage of SNPs of the SARS-CoV-2 genome, it is not practical to avoid all SNPs on every primer/probe binding site. Although false negative results may occur, many molecular tests tolerate a few single nucleotide mismatches, which have low or even no impact at all on their performance.

Table 1. Variants of SARS-CoV-2 genome and most common variants located on S-RBD.

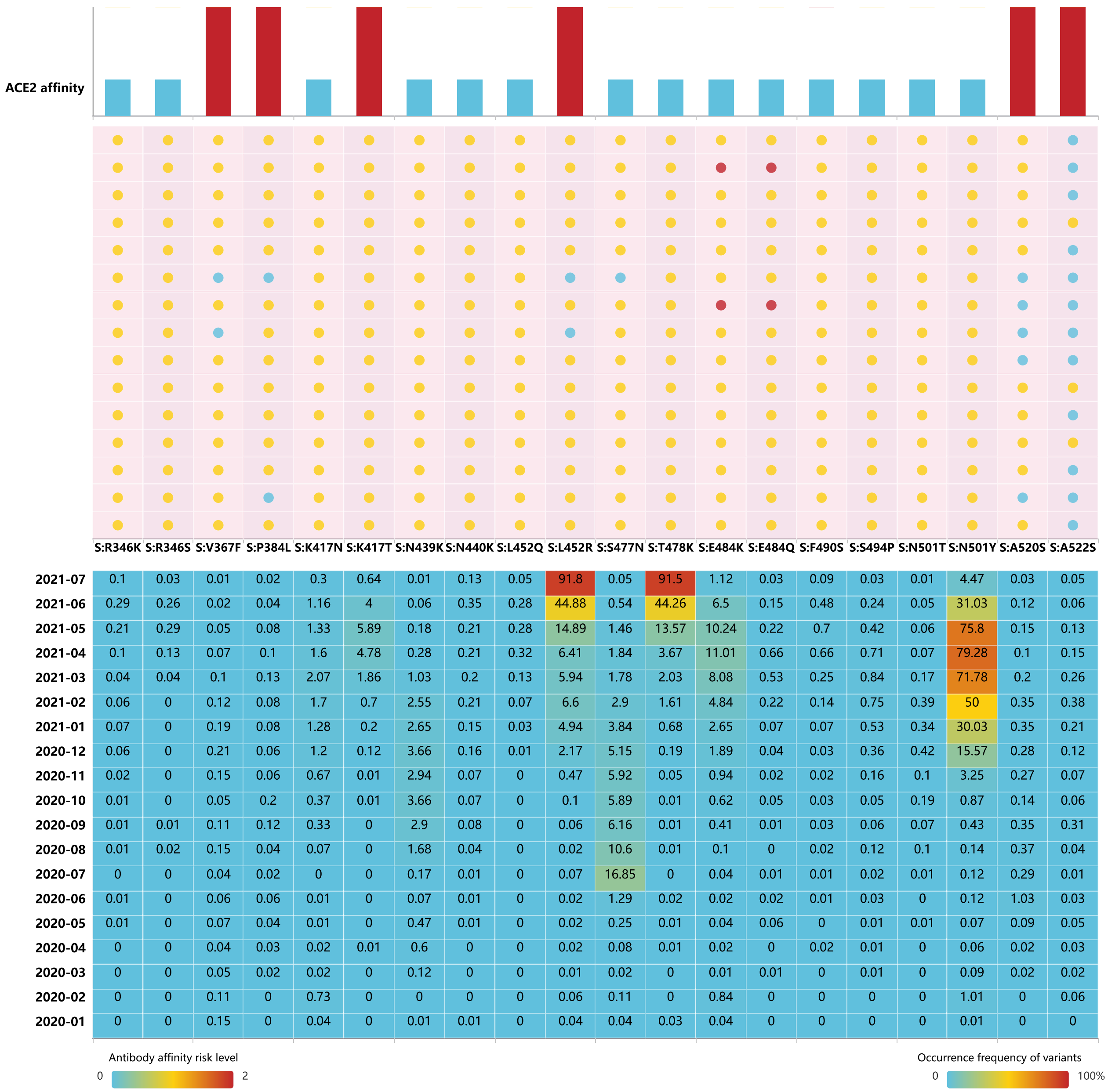
	Whole genome high frequency variants			RBD high frequency variants		
NO	Variants	Counts	Frequency	Variants	Counts	Frequency
1	S:D614G	2467291	95.85%	S:N501Y	1200001	46.62%
2	ORF1ab:P314L	2437459	94.69%	S:L452R	269897	10.49%
3	N:R203K	1434386	55.72%	S:T478K	206979	8.04%
4	N:G204R	1432232	55.64%	S:E484K	151017	5.87%
5	S:N501Y	1200001	46.62%	S:S477N	68895	2.68%
6	S:P681H	1178130	45.77%	S:K417T	57507	2.23%
7	ORF1ab:T1001I	1125623	43.73%	S:K417N	33585	1.30%
8	S:D1118H	1124839	43.70%	S:N439K	33447	1.30%
9	S:A570D	1122643	43.61%	S:S494P	12880	0.50%
10	S:T716I	1122555	43.61%	S:F490S	7757	0.30%
11	ORF8:Y73C	1120251	43.52%	S:E484Q	7179	0.28%
12	ORF1ab:A1708D	1118801	43.46%	S:A520S	5443	0.21%
13	N:S235F	1118673	43.46%	S:N440K	4610	0.18%
14	S:S982A	1116061	43.36%	S:A522S	4436	0.17%
15	ORF8:R52I	1113847	43.27%	S:N501T	4194	0.16%
16	N:D3L	1112519	43.22%	S:L452Q	3704	0.14%
17	ORF1ab:I2230T	1099897	42.73%	S:V367F	2499	0.10%
18	ORF3a:Q57H	456450	17.73%	S:R346K	2357	0.09%
19	ORF1ab:E265I	365975	14.22%	S:P384L	2253	0.09%
20	S:L452R	269897	10.49%	S:R346S	2188	0.09%

METHODS

Data sources and data processing

We extracted 2 635 714 SARS-CoV-2 sequences from the EpiCov™ section of the GISAID portal (10), and 956 676 SARS-CoV-2 sequences from the US National Center for Biotechnology Information (11) and NMDC (www.nmdc.cn). After low quality and duplicated sequences were removed, the final filtered raw sequence data set comprised 2 574 081 sequences. Each sequence was mapped against the reference genome from Wuhan, China (NCBI accession No. NC_045512.2) to identify mutations and deletions in the SARS-CoV-2 genome. We used the same site-numbering scheme as the reference genome to generate the lists of nucleotide variants and amino acids variants. Each mutation was then examined according to the following aspects (Figure 6):

- **Changes in free energy of binding with neutralizing antibodies caused by single amino acid mutation:** Saambe-3D (12) was used to predict changes in free energy of binding caused by single amino acid mutation and disruption of protein–protein interaction (PPI). Mutation types included destabilizing mutation ($\Delta\Delta G > 0$), stable mutation ($-1.5 < \Delta\Delta G < 0$), highly destabilizing mutation ($\Delta\Delta G > 1.5$) and highly stable mutation ($\Delta\Delta G < -1.5$). Subsequently, we predicted the affinity with 15 neutralizing antibodies (13–27), some of which have been approved as therapeutic antibodies for COVID-19 (casirivimab [28], imdevimab [28], bamlanivimab [29], etesevimab [29] and sotrovimab [30]). Finally, we assigned an overall ranked risk level from 1 to 3 based on the average $\Delta\Delta G$ values for all 15 antibodies.
- **Changes in free energy of binding with S protein and ACE2 induced by single amino acid mutation:** Saambe-3D was utilized to predict changes in free energy of binding caused by single amino acid mutation and whether that mutation could disrupt the PPI. Mutation types included destabilizing mutation ($\Delta\Delta G > 0$), stable mutation ($-1.5 < \Delta\Delta G < 0$), highly destabilizing mutation ($\Delta\Delta G > 1.5$) and highly stable mutation ($\Delta\Delta G < -1.5$). We assigned risk level 2 to highly stable mutations ($\Delta\Delta G < -1.5$) and risk level 1 to stable mutation ($-1.5 < \Delta\Delta G < 0$).
- **Difficulty of occurrence of nucleotide diversity:** This was represented by a 'nonsynonymous density' value reflecting the difficulty of the occurrence of nucleotide diversity and was evaluated by calculating the density of synonymous mutations and missense mutations under a sliding window. High-frequency variants that occurred before July 2021 were used as major alleles for statistical analysis. The density reflects the difficulty of occurrence of a mutation in a certain segment. High frequency densities indicate rapidly accumulating mutations in the region and low frequency densities may indicate a SNP desert (31), i.e. regions where potential selection of elimination occurs, implying that the virus has long-term and stable adaptive changes in this region.
- **Risk of replacement of amino acid:** PAM (32) and BLOSUM (33) matrices were employed to evaluate the risk of amino acid replacement. If replacement of two amino acids frequently occurred, it indicated that such amino acid replacements are stable. The replacement was assigned a low risk level and vice versa.
- **Effects of mutations on biological function of proteins:** 'Impact on protein function' was calculated using PROVEAN (34) to predict the effects of amino acid variants on the biological functions of proteins. The threshold for destructiveness and neutrality was set at −2.5.
- **Effect of variation on secondary structure:** Bepipred2.0 (35) was used for the 'secondary structure prediction' of the mutated protein and for comparison with the published X-ray diffraction data for the protein.
- **Effects of variation on potential continuous and discontinuous epitopes:** ElliPro (36) was used to predict 'changes of antigen continuous epitopes' and 'changes of antigen discontinuous epitopes' before and after the variation occurred.



	S:R346K	S:R346S	S:V367F	S:P384L	S:K417N	S:K417T	S:N439K	S:N440K	S:L452Q	S:L452R	S:S477N	S:T478K	S:E484K	S:E484Q	S:F490S	S:S494P	S:N501T	S:N501Y	S:A520S	S:A522S
2021-07	0.1	0.03	0.01	0.02	0.3	0.64	0.01	0.13	0.05	91.8	0.05	91.5	1.12	0.03	0.09	0.03	0.01	4.47	0.03	0.05
2021-06	0.29	0.26	0.02	0.04	1.16	4	0.06	0.35	0.28	44.88	0.54	44.26	6.5	0.15	0.48	0.24	0.05	31.03	0.12	0.06
2021-05	0.21	0.29	0.05	0.08	1.33	5.89	0.18	0.21	0.28	14.89	1.46	13.57	10.24	0.22	0.7	0.42	0.06	75.8	0.15	0.13
2021-04	0.1	0.13	0.07	0.1	1.6	4.78	0.28	0.21	0.32	6.41	1.84	3.67	11.01	0.66	0.66	0.71	0.07	79.28	0.1	0.15
2021-03	0.04	0.04	0.1	0.13	2.07	1.86	1.03	0.2	0.13	5.94	1.78	2.03	8.08	0.53	0.25	0.84	0.17	71.78	0.2	0.26
2021-02	0.06	0	0.12	0.08	1.7	0.7	2.55	0.21	0.07	6.6	2.9	1.61	4.84	0.22	0.14	0.75	0.39	50	0.35	0.38
2021-01	0.07	0	0.19	0.08	1.28	0.2	2.65	0.15	0.03	4.94	3.84	0.68	2.65	0.07	0.07	0.53	0.34	30.03	0.35	0.21
2020-12	0.06	0	0.21	0.06	1.2	0.12	3.66	0.16	0.01	2.17	5.15	0.19	1.89	0.04	0.03	0.36	0.42	15.57	0.28	0.12
2020-11	0.02	0	0.15	0.06	0.67	0.01	2.94	0.07	0	0.47	5.92	0.05	0.94	0.02	0.02	0.16	0.1	3.25	0.27	0.07
2020-10	0.01	0	0.05	0.2	0.37	0.01	3.66	0.07	0	0.1	5.89	0.01	0.62	0.05	0.03	0.05	0.19	0.87	0.14	0.06
2020-09	0.01	0.01	0.11	0.12	0.33	0	2.9	0.08	0	0.06	6.16	0.01	0.41	0.01	0.03	0.06	0.07	0.43	0.35	0.31
2020-08	0.01	0.02	0.15	0.04	0.07	0	1.68	0.04	0	0.02	10.6	0.01	0.1	0	0.02	0.12	0.1	0.14	0.37	0.04
2020-07	0	0	0.04	0.02	0	0	0.17	0.01	0	0.07	16.85	0	0.04	0.01	0.01	0.02	0.01	0.12	0.29	0.01
2020-06	0.01	0	0.06	0.06	0.01	0	0.07	0.01	0	0.02	1.29	0.02	0.02	0.02	0.01	0.03	0	0.12	1.03	0.03
2020-05	0.01	0	0.07	0.04	0.01	0	0.47	0.01	0	0.02	0.25	0.01	0.04	0.06	0	0.01	0.01	0.07	0.09	0.05
2020-04	0	0	0.04	0.03	0.02	0.01	0.6	0	0	0.02	0.08	0.01	0.02	0	0.02	0.01	0	0.06	0.02	0.03
2020-03	0	0	0.05	0.02	0.02	0	0.12	0	0	0.01	0.02	0	0.01	0.01	0	0.01	0	0.09	0.02	0.02
2020-02	0	0	0.11	0	0.73	0	0	0	0	0.06	0.11	0	0.84	0	0	0	0	1.01	0	0.06
2020-01	0	0	0.15	0	0.04	0	0.01	0.01	0	0.04	0.04	0.03	0.04	0	0	0	0	0.01	0	0

Figure 3. Binding stability to ACE2 and antibody affinity risk level for key mutations on S-RBD. Risk levels of reduced antibody affinity for 15 antibodies were calculated. The risk levels of antibody affinity and increased binding stability to ACE2 are ranked 0 to 2. Frequency of these variants over time are provided.

Effect of variation on effectiveness of detection reagents: For PCR 'Primer efficacy evaluation', the location and the frequency of the variant were considered comprehensively. If the variant occurred in the last three bases of the 3′ end, an early warning score will be given. In addition, the number of mutations was also assessed, and the corresponding score was given based on the number of variants at the last three bases in the 3′ end. The warning rating for RT-PCR primers was based on these scores.

Machine learning model for risk evaluation

We performed a comprehensive analysis of viral strain risk level by evaluating the difficulty of occurrence of nucleotide variants, possibility of amino acid replacement, change in protein secondary structure, and changes in ACE2 and neutralizing antibody free energy of binding caused by individual amino acid mutations. Each strain was given a series of characteristic quantities according to every mutation it carries. We constructed two random forest classifiers to verify the effectiveness of these characteristic quantities and used these parameters to group strains by their transmissibility and affinity with neutralizing antibodies.

The strains belong to eight WHO VOI/VOC were grouped into six groups according to two grouping modes: the normal transmissibility group, the mildly increased transmission group, the severely increased transmission

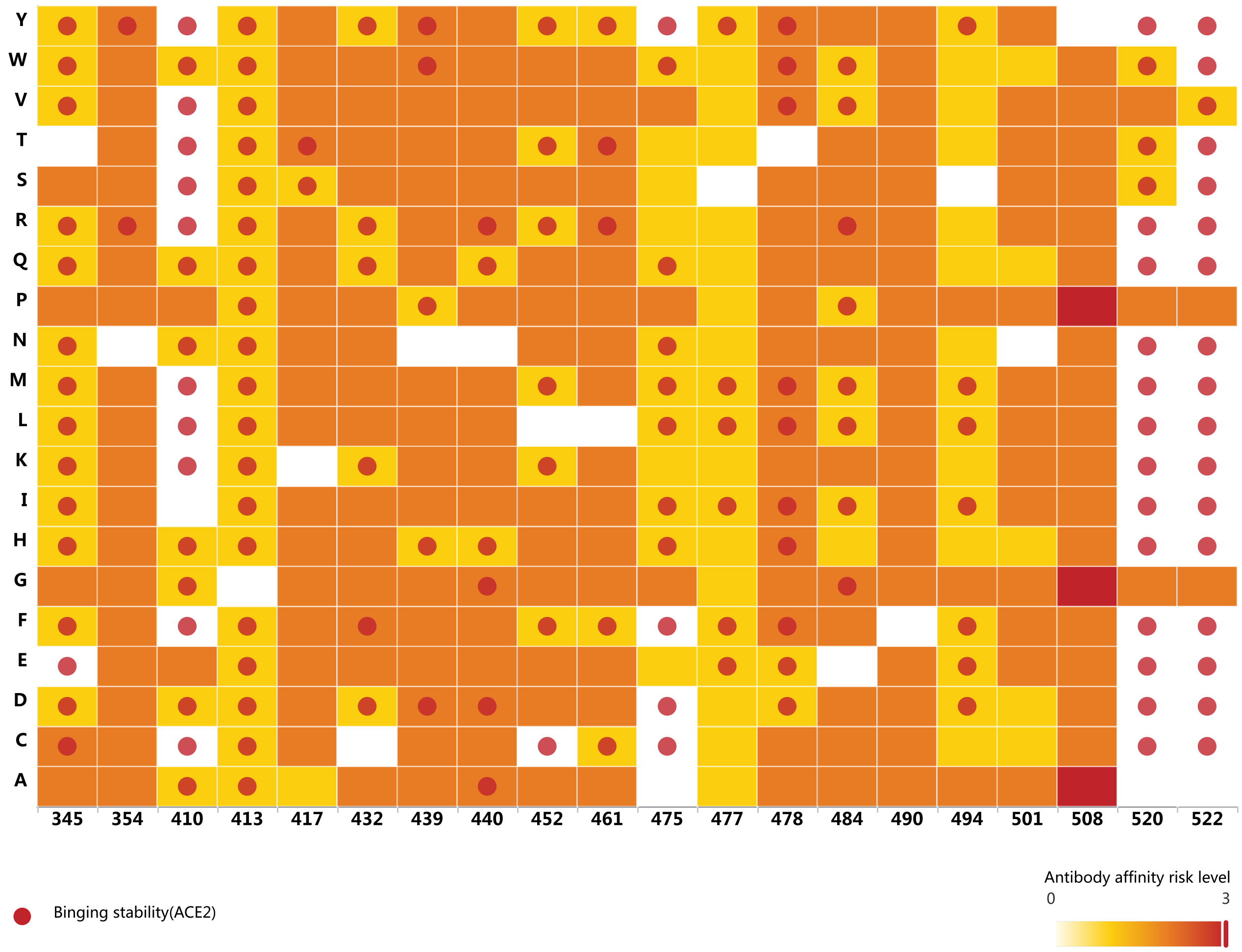


Figure 4. Binding stability to ACE2 and antibody affinity risk level for key known mutations and virtual mutations on S-RBD. A red dot indicates is increased binding stability to ACE2. Overall risk levels of reduced antibody affinity for 15 antibodies are ranked 0 to 3. Both known and virtual mutations were evaluated.

group, the normal affinity group, the mildly decreased affinity group and the severely decreased affinity group. Up to 50 000 complete genomic sequences were randomly extracted from the GISAID database for each of the eight VOI/VOC strains, and approximately 200 000 sequences were used to construct the model. All variant sites in the whole genome sequence of a strain were identified and parameters including the difficulty of occurrence of nucleotide variants, the possibility of amino acid replacement, the effect of variants on protein secondary structure, and changes in ACE2 and neutralizing antibody binding free energy caused by individual amino acid mutations were calculated for each variant site, which were then used to assign values to a strain sequence and construct the dataset. The Boruta algorithm was used to filter the feature measurements and the random forest algorithm was used to construct the classification model. To assess the reliability and stability of the model, 1000 random iterations were performed (70% were randomly selected as the training set and the remaining 30% as the testing set in each iteration). The prediction performance of the model was measured by area under the curve, accuracy, precision and sensitivity. Details of Machine Learning Model were provided in supplementary material (Supplementary Tables S1–S4 and Figures S1–S3).

CONCLUSION AND FUTURE DIRECTIONS

As of 5 August 2021, the number of confirmed COVID-19 patients worldwide reached 200 million with >4 million deaths. Over 70 vaccines are currently under development and 4 billion vaccine doses have already been administered (https://coronavirus.jhu.edu/map.html). Rapid diagnosis and vaccination are still the most effective methods for controlling the pandemic. As a result, it remains crucial to understand whether SARS-CoV-2 variants impact the affinity of current neutralizing antibodies under development or the performance of current diagnostic methods. It is also critical to pay close attention to variants that may escape from protective immune responses induced by population-level immunity. The VarEPS system presented here allows close monitoring and evaluation of the current global status of genetic variations of SARS-CoV-2.

VarEPS enables the user to focus on the updated global status of SARS-CoV-2 genome sequences and variation

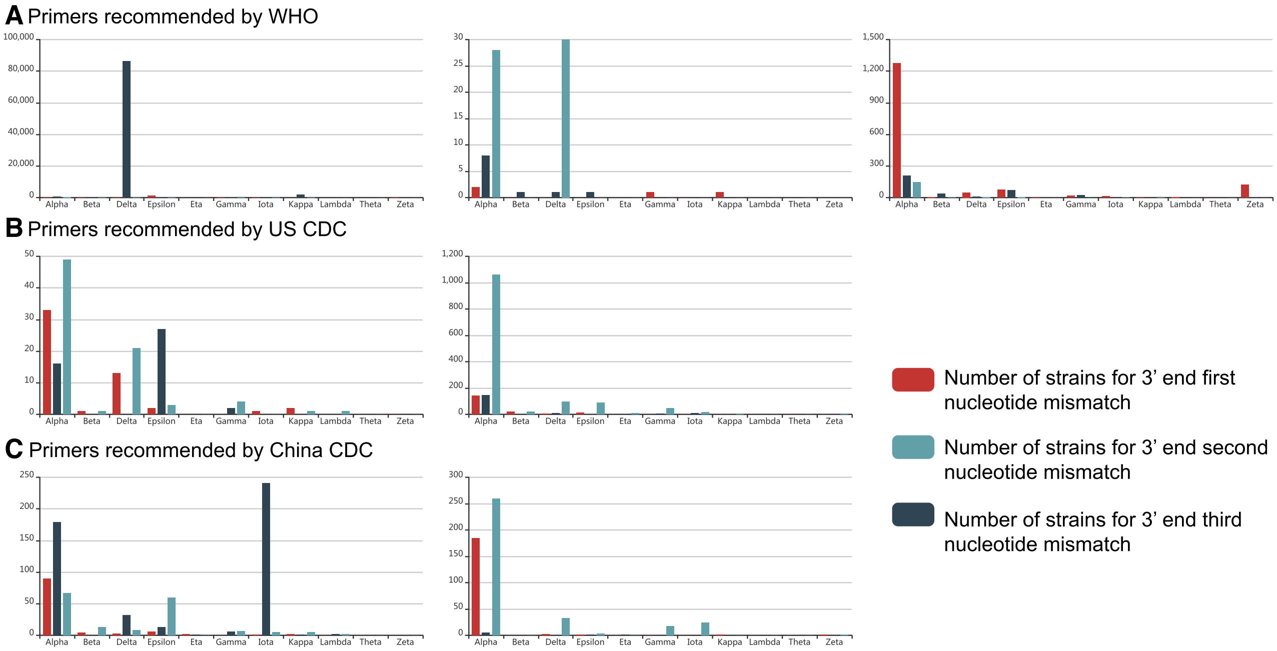


Figure 5. Nucleotide mismatch statistics for primers. Nucleotide mismatches were compared for the 3′ end of primers. The number of strains for each lineage were calculated.

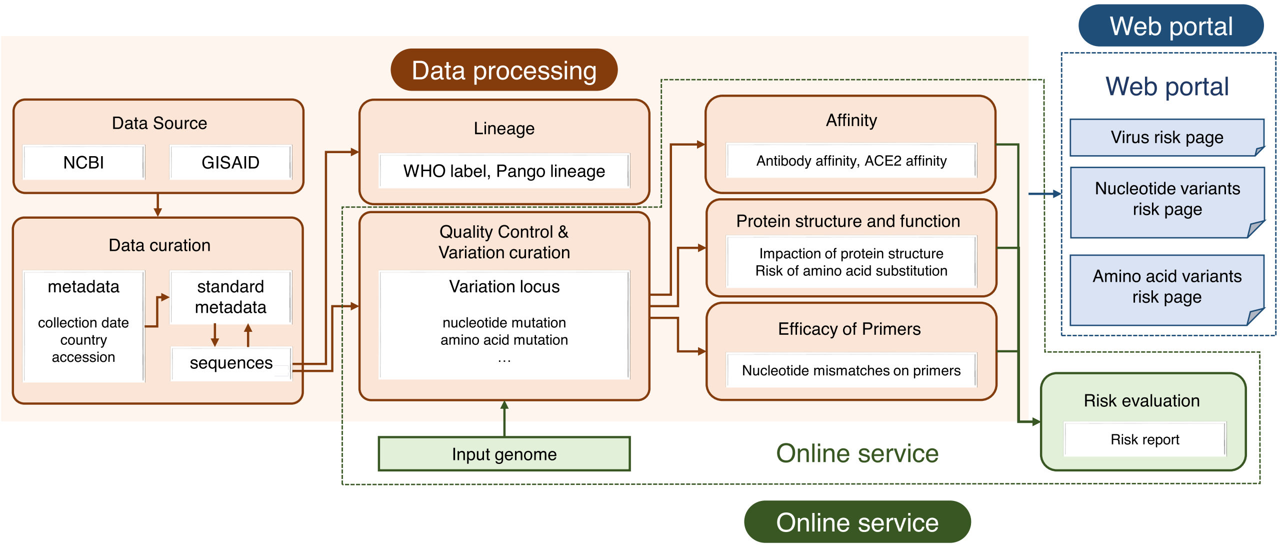


Figure 6. Schematic representation of VarEPS for data processing and online analysis service. SARS-CoV-2 genome sequences were integrated to perform metadata curation and quality control procedures. Sequence data were mapped to the reference genome for variation annotation. Each annotated variant was used to calculate effects on translation efficiency, secondary structure, binding capacity of ACE2 and neutralizing antibodies and efficacy of primers. Our web portal provides multiple query selections to display results on both known and virtual mutations. The system also provides online analysis service for custom submitted sequences.

analysis. It provides different levels of variant evaluation for translation efficiency, secondary structure, binding capacity of ACE2, binding capacity of neutralizing antibodies and efficacy of RT-PCR primers. Combined with the online analysis tools, the system can serve as both a navigation and recommendation tool for global virus variant surveillance. Moreover, the system can aid in designing robust vaccines and neutralizing monoclonal antibodies in the future. Based on the risk level evaluation of virtual variants, it provides key information for the design of prophylactic antibodies and vaccines that target variations with higher risk levels.

We will continuously update the system with new data on various resources of SARS-CoV-2 genome sequences. The machine learning model presented here is the first to successfully evaluate binding affinity and to group strains based on this attribute. The model will be further developed for broader evaluations. As more *in vitro* and *in vivo* studies are conducted, the *in silico* models will be iteratively optimized, and the simulation and prediction features will improve in accuracy with solid support from experimental results.

DATA AVAILABILITY

There are no access restrictions for academic use of the platform. Access to VarEPS is free at www.nmdc.cn/ncovn.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the GISAID Initiative and are grateful to all of the data contributors, including the authors, the originating laboratories responsible for obtaining the specimens, and the submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We would like to thank the National Pathogen Resource Center (NPRC), Chinese Center for Disease Control and Prevention for providing primer information. The numerical calculations in this paper have been done on CAS Xiandao-1 computing environment.

FUNDING

National Key Research Program of China [2021YFC0863300]; Chinese Academy of Sciences [XDA19050301, WX145XQ07-01, KFZD-SW-219]; National Science Foundation [82161148010]. Funding for open access charge: National Key Research Program of China [2021YFC0863300].
Conflict of interest statement. None declared.

REFERENCES

1. Wu,A.P., Wang,L.I., Zhou,H.Y., Ji,C.Y., Xia,S.Z., Cao,Y., Meng,J., Ding,X., Gold,S., Jiang,T.J. *et al.* (2021) One year of SARS-CoV-2 evolution. *Cell Host Microbe*, **29**, 503–507.
2. van Dorp,L., Acman,M., Richard,D., Shaw,L.P., Ford,C.E., Ormond,L., Owen,C.J., Pang,J., Tan,C.C.S., Boshier,F.A.T. *et al.* (2020) Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection. *Genetics Evol.*, **83**, 104351.
3. LaTourrette,K., Holste,N.M., Rodriguez-Peña,R., Leme,R.A. and Garcia-Ruiz,H. (2021) Genomewide variation in betacoronaviruses. *Virol*, **95**, e00496–21.
4. Yurkovetskiy,L., Wang,X., Pascal,K.E., Tomkins-Tinch,C., Nyalile,T., Wang,Y.T., Baum,A., Diehl,W.E., Dauphin,A., Carbone,C. *et al.* (2020) Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell*, **3**, 739–751.
5. Laurini,E., Marson,D., Aulic,S., Fermeglia,A. and Pricl,S. (2021) Computational mutagenesis at the SARS-CoV-2 spike protein/angiotensin-converting enzyme 2 binding interface: comparison with experimental evidence. *ACS Nano*, **15**, 6929−6948
6. Jiang,C.S., Li,X.W., Ge,C.R., Ding,Y.Y., Zhang,T., Cao,S., Meng,L.S. and Lu,S.M. (2021) Molecular detection of SARS-CoV-2 being challenged by virus variation and asymptomatic infection. *J. Pharm. Analysis*, **11**, 257–264.
7. Marcais,G., Delcher,A.L., Phillippy,A.M., Coston,R., Salzberg,S.L. and Zimin,A. (2018) MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.*, **14**, e1005944.
8. McCallum,M., Bassi,J., Marco,D.A., Chen,A., Walls,A.C., Iulio,J.D., Tortorici,M. A., Navarro,M.J., SilacciFregni,C., Saliba,C. *et al.* (2021) SARS-CoV-2 immune evasion by the B.1.427/B.1.429 variant of concern. *Science*, **373**, 648–654.
9. Chen,J.H., Wang,R., Wang,M.L. and Wei,G.W. (2020) Mutations strengthened SARS-CoV-2 infectivity. *J. Mol. Biol.*, **432**, 5212–5226.
10. Shu,Y. and McCauley,J. (2017) GISAID: from vision to reality. *EuroSurveillance*, **22**, 30494.
11. Sayers,E.W., Beck,J., Bolton,E.E., Bourexis,D., Brister,J.R., Canese,K., Comeau,D.C., Funk,K., Kim,S., Klimke,W. *et al.* (2021) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **49**, D10–D17.
12. Pahari,S., Li,G., Murthy,A.K., Liang,S.Q., Fragoza,R., Yu,H.Y. and Alexov,E. (2020) SAAMBE-3D: predicting effect of mutations on protein–protein interactions. *Int. J. Mol. Sci.*, **21**, 2563.
13. Lv,Z., Deng,Y.Q., Ye,Q., Cao,L., Sun,C.Y., Fan,C., Huang,W., Sun,S., Sun,Y., Zhu,L. *et al.* (2020) Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic antibody. *Science*, **369**, 1505–1509.
14. Hansen,J., Baum,A., Pascal,K.E., Russo,V., Giordano,S., Wloga,E., Fulton,B.O., Yan,Y., Koon,K., Patel,K. *et al.* (2020) Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science*, **369**, 1010–1014.
15. Kreye,J., Reincke,S.M., Kornau,H.C., Sanchez-Sendin,E., Corman,V.M., Liu,H., Yuan,M., Wu,N.C., Zhu,X., Lee,C.D. *et al.* (2020) A therapeutic Non-self-reactive SARS-CoV-2 antibody protects from lung pathology in a COVID-19 hamster model. *Cell*, **183**, 1058.
16. Huo,J., Zhao,Y., Ren,J., Zhou,D., Duyvesteyn,H.M.E., Ginn,H.M., Carrique,L., Malinauskas,T., Ruza,R.R., Shah,P.N.M. *et al.* (2020) Neutralization of SARS-CoV-2 by destruction of the prefusion spike. *Cell Host Microbe*, **28**, 445–454.
17. Zhou,D., Duyvesteyn,H.M.E., Chen,C.P., Huang,C.G., Chen,T.H., Shih,S.R., Lin,Y.C., Cheng,C.Y., Cheng,S.H., Huang,Y.C. *et al.* (2020) Structural basis for the neutralization of SARS-CoV-2 by an antibody from a convalescent patient. *Nat. Struct. Mol. Biol.*, **27**, 950–958.
18. Barnes,C.O., Jette,C.A., Abernathy,M.E., Dam,K.A., Esswein,S.R., Gristick,H.B., Malyutin,A.G., Sharaf,N.G., Huey-Tubman,K.E., Lee,Y.E. *et al.* (2020) SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature*, **588**, 682–687.
19. Piccoli,L., Park,Y.J., Tortorici,M.A., Czudnochowski,N., Walls,A.C., Beltramello,M., Silacci-Fregni,C., Pinto,D., Rosen,L.E., Bowen,J.E. *et al.* (2020) Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell*, **183**, 1024–1042.
20. Supasa,P., Zhou,D., Dejnirattisai,W., Liu,C., Mentzer,A.J., Ginn,H.M., Zhao,Y., Duyvesteyn,H.M.E., Nutalai,R., Tuekprakhon,A. *et al.* (2021) Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. *Cell*, **184**, 2201.
21. Shi,R., Shan,C., Duan,X., Chen,Z., Liu,P., Song,J., Song,T., Bi,X., Han,C., Wu,L. *et al.* (2020) A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature*, **584**, 120–124
22. Kim,C., Ryu,D.K., Lee,J., Kim,Y.I., Seo,J.M., Kim,Y.G., Jeong,J.H., Kim,M., Kim,J.I., Kim,P. *et al.* (2021) A therapeutic neutralizing antibody targeting receptor binding domain of SARS-CoV-2 spike protein. *Nat. Commun.*, **12**, 288–288.
23. Jones,B.E., Brown-Augsburger,P.L., Corbett,K.S., Westendorf,K., Davies,J., Cujec,T.P., Wiethoff,C.M., Blackbourne,J.L., Heinz,B.A., Foster,D. *et al.* (2021) The neutralizing antibody, LY-CoV555, protects against SARS-CoV-2 infection in nonhuman primates. *Sci. Transl. Med.*, **13**, eabf1906.
24. Starr,T.N., Czudnochowski,N., Liu,Z., Zatta,F., Park,Y.J., Addetia,A., Pinto,D., Beltramello,M., Hernandez,P., Greaney,A.J. *et al.* (2021) SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. *Nature*, **597**, 97–102.
25. Bravo,J.P.K., Dangerfield,T.L., Taylor,D.W. and Johnson,K.A. (2021) Remdesivir is a delayed translocation inhibitor of SARS-CoV-2 replication. *Mol. Cell*, **81**, 1548–1552.
26. Huo,J., Le Bas,A., Ruza,R.R., Duyvesteyn,H.M.E., Mikolajek,H., Malinauskas,T., Tan,T.K., Rijal,P., Dumoux,M., Ward,P.N. *et al.* (2020) Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2. *Nat. Struct. Mol. Biol.*, **27**, 846–854.

27. Ge,J.W., Wang,R.K., Ju,B., Zhang,Q., Sun,J., Chen,P., Zhang,S.Y., Tian,Y.L., Shan,S.S. and Cheng,L. (2021) Antibody neutralization of SARS-CoV-2 through ACE2 receptor mimicry. *Nat. Commun.*, **12**, 250.
28. Phan,A.T., Gukasyan,J., Arabian,S., Wang,S. and Neeki,M.M. (2021) Emergent inpatient administration of casirivimab and imdevimab antibody cocktail for the treatment of COVID-19 pneumonia. *Cureus*, **13**, e15280.
29. (2021) An EUA for bamlanivimab and etesevimab for COVID-19. *Med. Lett. Drugs Ther.*, **63**, 49–50.
30. (2021) An EUA for sotrovimab for treatment of COVID-19. *Med. Lett. Drugs Ther.*, **63**, 97–98.
31. Wang,L., Hao,L., Li,X., Hu,S., Ge,S. and Yu,J. (2009) SNP deserts of Asian cultivated rice: genomic regions under domestication. *J. Evol. Biol.*, **22**, 751–761.
32. Wilbur,W.J. (1985) On the PAM matrix model of protein evolution. *Mol. Biol. Evol.*, **2**, 434–447.
33. Henikoff,S. and Henikoff,J.G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins*, **17**, 49–61.
34. Choi,Y. and Chan,A.P. (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, **31**, 2745–2747.
35. Jespersen,M.C., Peters,B., Nielsen,M. and Marcatili,P. (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic. Acids. Res.*, **45**, W24–W29.
36. Ponomarenko,J., Bui,H.H., Li,W., Fusseder,N., Bourne,P.E., Sette,A. and Peters,B. (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BioMed Central*, **9**, 514.