
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Biosafety and Biosecurity 4 (2022) 114–120

Contents lists available at ScienceDirect

Journal of Biosafety and Biosecurity

journal homepage: www.keaipublishing.com/en/journals/journal-of-biosafety-and-biosecurity/

Research Article

# "Mutation blacklist" and "mutation whitelist" of SARS-CoV-2

Yamin Sun [a,b,*,1], Min Wang [b,d,1], Wenchao Lin [b], Wei Dong [b], Jianguo Xu [a,c,e,*]

[a] Research Institute of Public Health, Nankai University, Tianjin, PR China
[b] Research Center for Functional Genomics and Biochip, Tianjin, PR China
[c] State Key Laboratory for Infectious Disease Prevention and Control, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 202206, PR China
[d] TEDA Institute of Biological Sciences and Biotechnology, Nankai University, PR China
[e] Research Units of Discovery of Unknown Bacteria and Function, Chinese Academy of Medical Sciences, Beijing 100730, PR China

ARTICLE INFO

*Article history:*
Received 31 May 2022
Received in revised form 21 June 2022
Accepted 27 June 2022

*Keywords:*
SARS-CoV-2
Mutation saturation
*De novo* mutations
Transmission

ABSTRACT

Over the past two years, scientists throughout the world have completed more than 6 million SARS-CoV-2 genome sequences. Today, the number of SARS-CoV-2 genomes exceeds the total number of all other viral genomes. These genomes are a record of the evolution of SARS-CoV-2 in the human host, and provide information on the emergence of mutations. In this study, analysis of these sequenced genomes identified 296,728 *de novo* mutations (DNMs), and found that six types of base substitutions reached saturation in the sequenced genome population. Based on this analysis, a "mutation blacklist" of SARS-CoV-2 was compiled. The loci on the "mutation blacklist" are highly conserved, and these mutations likely have detrimental effects on virus survival, replication, and transmission. This information is valuable for SARS-CoV-2 research on gene function, vaccine design, and drug development. Through association analysis of DNMs and viral transmission rates, we identified 185 DNMs that positively correlated with the SARS-CoV-2 transmission rate, and these DNMs where classified as the "mutation whitelist" of SARS-CoV-2. The mutations on the "mutation whitelist" are beneficial for SARS-CoV-2 transmission and could therefore be used to evaluate the transmissibility of new variants. The occurrence of mutations and the evolution of viruses are dynamic processes. To more effectively monitor the mutations and variants of SARS-CoV-2, we built a SARS-CoV-2 mutation and variant monitoring and pre-warning system (MVMPS), which can monitor the occurrence and development of mutations and variants of SARS-CoV-2, as well as provide pre-warning for the prevention and control of SARS-CoV-2 (https://www.omicx.cn/). Additionally, this system could be used in real-time to update the "mutation whitelist" and "mutation blacklist" of SARS-CoV-2.



## 1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of the ongoing coronavirus 2019 (COVID-19) pandemic, belongs to the *Sarbecovirus* genus in the *Coronaviridae* family.[1,2] The first outbreak cases of SARS-CoV-2 were detected in December 2019 and quickly spread globally,[3–6] leading to the World Health Organization (WHO) assigning the virus pandemic status in March 2020. As of February 11, 2022, 404,910,528 cases of SARS-CoV-2 had been confirmed, resulting in 5,783,776 deaths reported to the WHO.[7] As of January 18, 2022, more than 6 million genomes had been reported,[8] which exceeded the total number of genomes for all other viruses. These genome sequences fully record the evolution of SARS-CoV-2 during the pandemic, and provide important mutation information.

Since the COVID-19 pandemic first began, SARS-CoV-2 has continuously evolved with many variants emerging across the world.[9] These variants are categorized as variants of interest (VOI), variants of concern (VOC), and variants under monitoring (VUM) based on their transmission potential. As of February 2022, there were five SARS-CoV-2 lineages designated as VOC (Alpha, Beta, Gama, Delta, and Omicron). VOC have increased transmissibility compared with the original virus and the potential for increased disease severity.[10,11] In addition, VOC exhibit decreased susceptibility to vaccine-induced or infection-induced immunity, and thus possess the ability to re-infect previously infected and recovered individuals.

* Corresponding authors at: Research Institute of Public Health, Nankai University, Tianjin, PR China.
*E-mail addresses:* nksunyamin@aliyun.com (Y. Sun), xujianguo@icdc.cn (J. Xu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jobb.2022.06.006
2588-9338/

A mutation is defined as an alteration in the DNA or RNA sequence of a genome, which consequently confers a new genotype and sometimes a new phenotype. Mutations can be beneficial, neutral, or harmful for the virus.[12–14] Beneficial mutations may help the virus spread or replicate more efficiently, providing an advantage over other strains.[15,16] Harmful mutations affect virus replication and transmission and will not be retained and recorded.[17,18] This evolutionary process guides SARS-CoV-2 adaptation to its new host.[19] A large number of SARS-CoV-2 mutations have been recorded in more than 6 million genomes, raising the question of which mutations are beneficial or deleterious for SARS-CoV-2. Since the rapid emergence of mutations in viral RNA could potentially render vaccines ineffective, drug therapy unsuccessful, and lead to false detection data, it is critical to analyze and understand the implications of SARS-Cov-2 single-nucleotide mutations.[20,21]

In this study, analysis of SARS-CoV-2 mutations identified six types of base substitutions that reached saturation in the population. Deleterious mutations with a negative impact on virus survival, replication, and transmission were identified and classified as the "mutation blacklist". Beneficial mutations associated with an increased transmission rate were identified and classified as the "mutation whitelist". The "mutation blacklist" and "mutation whitelist" determined for SARS-CoV-2 in this study are of great value for research on gene function, vaccine design, drug development, and the identification of new VOC.

## 2. Materials and methods

### 2.1. Data collection

SARS-CoV-2 sequences were retrieved from the Global Initiative on Sharing Avian Influenza Data (GISAID) initiative database (as of 18 January 2022, https://www.gisaid.org).[22,23] Complete genomes with an N-content lower than 0.01% and high coverage were selected for subsequent analysis. A Multiple Alignment using Fast Fourier Transform (MAFFT)-generated alignment of high coverage complete genome sequences was downloaded from the website.

### 2.2. Mutation analysis

The complete genome of the SARS-CoV-2 isolate Wuhan-Hu-1 (NC_045512.2) was used as the reference genome; mutations in all other samples were compared to this reference isolate. Detected mutations were confirmed using Integrative Genomics Viewer (IGV)[24] and annotated with the SnpEff program.[25]

### 2.3. Construction of a phylogenetic tree

Construction of the phylogenetic tree was performed as previously described.[26] The amount of computation needed to construct an evolutionary tree for the 2.8 million genomes is substantial. Hence, to improve computational efficiency, SARS-CoV-2 genomes were classified by pangolin lineages using the pangoLEARN algorithm.[27] The 2.8 million genomes were divided into 1,514 subsets according to their pangolin lineage. The RAxML[24] software was used to determine the topological relationship between each subset according to their common mutations, and to construct the evolutionary tree as a "root-tree". The maximum likelihood phylogenetic tree was constructed based on the General Time Reversible + Invariant + gamma sites (GTR + I + G) model of nucleotide substitution with 1000 bootstrap replicates. Then, 1,514 evolutionary trees were constructed as "branch-trees" for the 1,514 subset trees using the FastTree[25] software with the Jukes–Cantor model. Finally, "root-tree" and "branch-trees" were merged to generate the "final-tree" by an in-house script. The flowchart of evolutionary tree construction is provided in the supplementary information (Fig. S1).

### 2.4. De novo mutation detection

Construction of the phylogenetic tree was performed as previously described.[26] The information on the distribution of each mutation in the different clades of the "final-tree" was determined using an in-house-developed script. For each mutation, we step-by-step scanned the "final-tree" from root to tip to determine the proportion of mutations in each clade. When >50% of the genomes in a clade contained a particular mutation, we assumed the ancestor node of the clade contained the *de novo* mutations (DNMs). To avoid the identification of inherited mutations as DNMs by inaccurate terminal branching, we merged the DNMs that satisfied all of the following conditions: (1) share the same mutation type, such as C10029T (base position 10,029 in the genome is mutated from C to T), (2) appear in the same clade and the clade size is < 2,000 genomes, (3) isolated from the same country, and (4) had a time span of < 6 months. We used these criteria because of the very low probability of detecting multiple DNMs in the same country among 2000 genomes within 6 months. If the mutation rate for SARS-CoV-2 is calculated as $3 \times 10^{-3}$ nucleotide substitutions per site per year, the probability of detecting the same DNMs in 2000 genomes within 6 months should be: p = 0.009 ($3 \times 10^{-3} * 3 \times 10^{-3} * 2000 * 0.5$). To avoid the impact of sequencing errors on DNM detection, we filtered out DNMs based on a single genome. The flowchart of DNM detection is provided in the supplementary information (Fig. S2).

### 2.5. Mutation saturation

To analyze whether all 12 base substitutions (A->T, A->C, A->G, C->A, C->T, C->G, T->A, T->C, T->G, G->A, G->T, and G->C) in SARS-CoV-2 are saturated, we first grouped the DNMs into 12 subsets according to the base substitution type, and then sorted them according to the time of occurrence. We counted the number of newly occurring non-redundant DNMs week by week, and used R scripts to draw saturation curves of 12 base substitution types. Mutation saturation was defined as the timepoint when the number of non-redundant DNMs of a base substitution type stagnates, indicative of a plateau in the saturation curve.

### 2.6. Association between DNMs and the transmission rate

Positive effects in transmission rate were estimated from the increased genomic prevalence of a specific DNM in the subsequently sequenced genomes. For each DNM, the proportion of genomes containing this mutation, out of all genomes sequenced in the 10 weeks following its emergence, was calculated. These data were then analyzed by linear least-squares regression using SciPy[28] to derive the proportion growth slope of each DNM. The slope value of each DNM represents its influence on the transmission potential of each SARS-CoV-2 variant, with larger values reflecting a greater positive impact on the viral transmission rate.

## 3. Results and discussion

### 3.1. Mutation saturation

DNM is a term used in genetics to describe a type of genetic mutation that develops in a family member for the first time. Virus evolution begins with a DNM in the viral genome. In this study, we detected a total of 297,826 DNMs in SARS-CoV-2, which covered

88% of the total genome of the virus. This raises the question of why the remaining 12% of the genome could not be identified. Theoretically, it could be random if the number of sequenced genomes is too short to provide full coverage, or alternatively because those sites result in fatal mutations. To address this non-exclusive hypothesis, we analyzed the mutation saturation of 12 types of base substitution (A->T, A->C, A->G, C->A, C->T, C->G, T->A, T->C, T->G, G->A, G->T, and G->C). Previous studies showed that different base substitutions in SARS-CoV-2 had different mutation rates; e.g., G->T and C->T had a high mutation rate, while A->T, T->A, C->G and G->C had a low mutation rate. Our analysis showed that six base substitution types, C->T, G->T, G->A, T->C, A->G, and A->C, reached saturation by January 2022. Mutation saturation means that no novel mutations should appear in the future, even if more genomes are sequenced. Among them, the base substitution type G->T reached saturation as early as December 27, 2020, and the base substitution type C->T reached saturation on February 21, 2021, while G->A, C->A, T->C, and A->G reached saturation on April 4, 2021, December 19, 2021, November 14, 2021, and October 10, 2021, respectively (Fig. 1). The other six base substitution types (C->G, G->C, A->T, T->A, T->G, and A->C) had not reached mutation saturation as of February 2022. These observations are consistent with a previous report showing that these six base substitution types have a lower mutation frequency compared with other substitution types.[8,29] The mutation saturation of the six types of base substitution means that the subsequent variants in SARS-CoV-2 tend to be a combination of the existing high-frequency DNMs, rather than the emergence of a new mutation. To avoid the effect of sequencing errors on mutation saturation, complete genomes with an N-content lower than 0.01% and high coverage were selected for subsequent analysis. In addition, we filtered out DNMs supported by a single genome. Therefore, we believe that sequencing error had little effect on the mutation saturation.

### 3.2. "Mutation blacklist" of SARS-CoV-2

Among the six saturated base substitution types, we identified 5,945 potential nucleotide mutations located on 4,178 loci of all sequenced SARS-CoV-2 genomes that never mutated. Of these 5,945 potential nucleotide mutations, 1,039 were in protein coding regions and resulted in premature termination codons and consequent protein truncation, which explains the fatal phenotype. Excluding these, 4,906 potential nucleotide mutations at 3,308 loci never occurred in all SARS-CoV-2 sequenced genomes. These mutations may reflect significant changes in gene function with deleterious effects on viral survival, replication, and transmission. Therefore, these were designated as the "mutation blacklist" (Table S1). On the "mutation blacklist", 19 mutations were located in non-coding regions, accounting for 0.41%, and the remaining 4,887 mutations were located in coding regions, accounting for 99.59%.

Among the 4,887 mutations located in the coding region, 4,863 mutations (99.51%) were non-synonymous mutations, and a total of 3,778 mutations resulted in more hydrophilic amino acid residues, accounting for 77.31% of all non-synonymous mutations. This change in hydrophilicity may affect the structure of the corresponding protein and consequently gene function. Meanwhile, this result also implied that SARS-CoV-2 is evolving toward avoiding hydrophilic amino acids. On the "mutation blacklist", we found that mutations corresponding to 172 amino acid residues have never been changed. These amino acids may play an important role in maintaining corresponding protein structure and gene function. For example, in the receptor-binding domain (RBD) region of the spike protein, D398 is never mutated in all sequenced genomes. Previous studies showed that D398 contributes to variations in pH-dependence between the locked, closed, and open forms of the RBD. Thus, these unchanged 172 amino acids on various loci on the "mutations blacklist" may reveal functionally and structurally important amino acid residues.

Unlike mutations in coding regions, mutations in non-coding regions do not affect gene function or amino acids residues, and are generally subject to high mutation rates. In this study, we found highly conserved loci in non-coding regions of the genome. This means that these nucleotides may encode functional non-coding RNA or enzyme recognition sites. Analysis of the chromosomal distribution of the "mutation blacklist" identified a highly con-

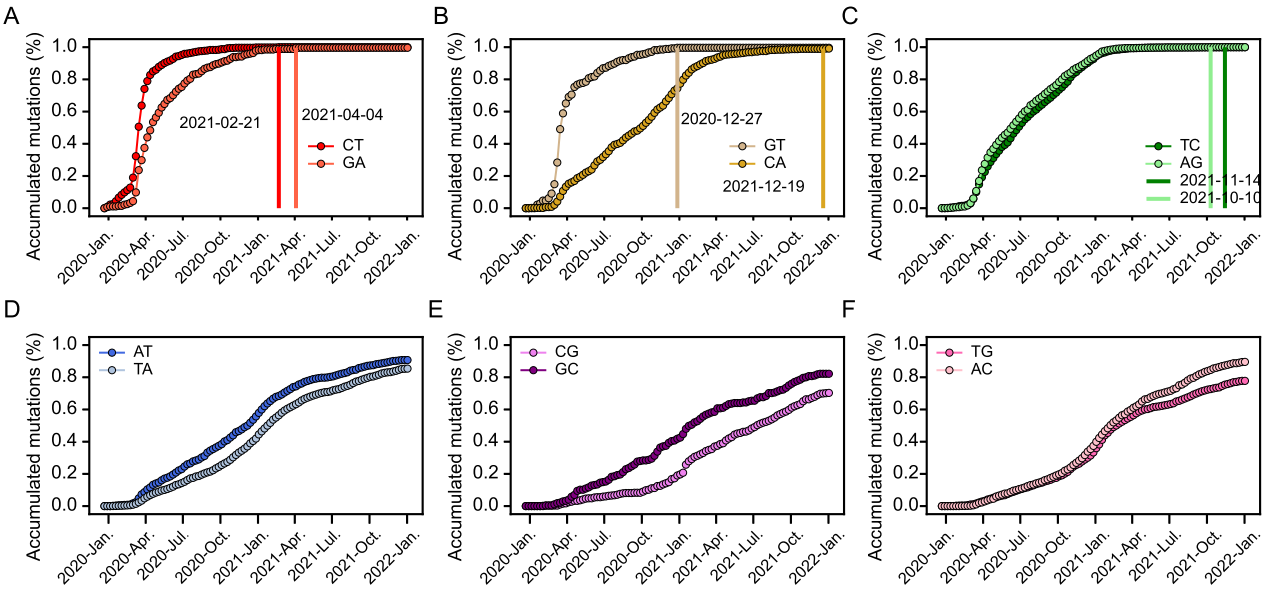


**Fig. 1.** Mutation saturation curves of 12 types of base substitution. Identified DNMs in the SARS-CoV-2 sequenced genomes were grouped by base type and the cumulative saturation percentage of C-> T and G-A (A), C->A and G->T (B), T->C and A->G (C), A->T and T->A (D), C->G and G->C (E), T->G and A->C (F), and were plotted over time. The curve plateau reflects 100% saturation, which is marked in the figure by a vertical line.

served region between 128 and 135 bp at the 5′UTR of the genome. Further analysis of this region identified a sequence motif (5′-TATAATTA-3′ motif), which was similar to a TATA box and may therefore be important for transcriptional regulation (Fig. 2A). Another example was the 5′-ACGAAC-3′ motif in the intergenic region. Non-mutated nucleotides are also enriched in this motif and previous studies[30] demonstrated that ACGAAC is a core transcription-regulating sequence guiding the discontinuous RNA synthesis of SARS-CoV-2 (Fig. 2B).

Based on the above analysis, we believe that the "mutation blacklist" identifies important amino acid residues in the coding region and the regulatory non-coding regions of the genome. These sites on the "mutation blacklist" are potential targets for the development of SARS-CoV-2 drugs or broadly reactive antibodies.

### 3.3. "Mutation whitelist" of SARS-CoV-2

Some mutations may affect the virus transmission rate. For example, since the emergence of the "star mutation" D614G, the proportion of SARS-CoV-2 variants with this mutation increased rapidly in the sequenced population, reflecting its high transmission rate.[31] Hence, we next determined the relationship between DNMs and variant transmission rates by tracking the changes in prevalence of each DNM in all genomes sequenced up to 10 weeks after its emergence. If the prevalence of the mutation increased weekly, then it is likely to be positively related to virus transmission, which can be measured by the slope of the corresponding linear regression model. Of all DNMs, a total of 185 mutations were significantly positively correlated with viral transmission (slope > 0.001). The chromosomal distribution of these mutations showed that most were concentrated in the RBD regions of SARS-CoV-2 Spike protein, M protein, and N protein, with a few located within the *orf1ab* gene and other genes encoding accessory proteins. Moreover, these mutation sites cover almost all of the important mutation sites of VOC and VOI (Fig. 3, Table S2). Therefore, these mutations were designated as the "mutation whitelist" that may benefit virus transmission. It should be noted that some of these mutations are "driver" mutations with a real biological impact on the viral transmission rate, while others may be "passenger" mutations resulting from a "free-riding tendency". The method used in this study could not distinguish "driver" and "passenger" mutations.

### 3.4. Mutation and variant monitoring and the pre-warning system

The occurrence of mutations and the evolution of viruses are dynamic processes. Consequently, the "mutation whitelist" and "mutation blacklist" of SARS-CoV-2 will change with the evolution of the virus. To monitor the mutations and the variants of SARS-CoV-2, we built a website (https://www.omicx.cn/) that presents a dynamic curve in real-time with the proportions of each mutation and variant among all SARS-CoV-2 strains. Through this real-time dynamic curve, we will be able to monitor the epidemic trend for mutations and variants. It should be noted that it can take ~2 weeks from sampling to submitting the sequenced genome data into public databases. Therefore, there will be a 2-week delay in the information detected for this website. This is a universal problem for all monitoring systems based on sequence data. The system consists of four functional modules with the following capabilities: (1) the system can collect SARS-CoV-2 published data resources from public databases (GISAID and NCBI) in real-time; (2) the system can calculate real-time statistics on the proportion of mutations and variants worldwide and for specific countries; (3) the system can assign each mutation and variant a growth rate according to the changing trend for the mutation and variant worldwide and for specific countries, and can estimate the possibility of each mutation and variant becoming a major epidemic strain in the world and various countries according to the growth

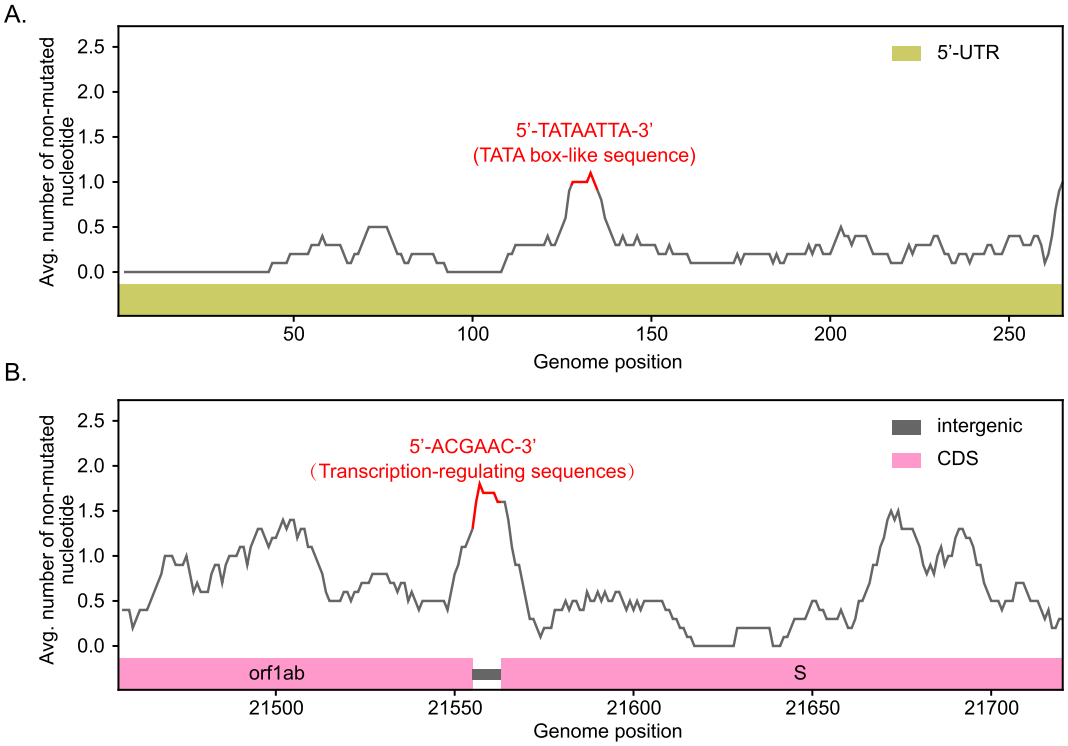


**Fig. 2.** A Sequence conservation analysis in the 5′-UTR. The x-axis is the position on the genome, and the y-axis is the number of mutations that have never been detected. B is the conservative analysis of the sequence in the intergenic region of the *orf1ab* and *spike* genes. The x-axis is the position on the genome, and the y-axis is the number of mutations that have never been detected.

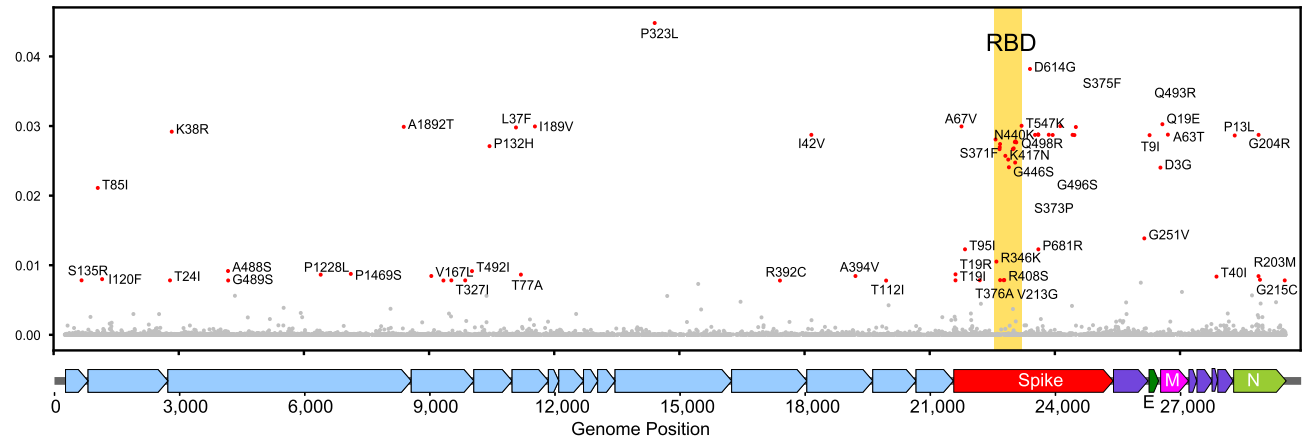


**Fig. 3.** Association between DNMs and transmission rate. The x-axis represents the SARS-CoV-2 genome position, and the y-axis represents the weekly growth slope of each DNM in the SARS-CoV-2 population. Mutations with a growth slope greater than 0.001 are marked in red and the corresponding amino acid mutations are noted.

rate; (4) the system can update the "mutation blacklist" and "mutation whitelist" of SARS-CoV-2 in real-time, and highlight the mutation sites that have appeared for the first time. Three data categories were included in this website: variant monitoring, mutation monitoring, and mutation blacklist and whitelist (Fig. 4A).

(1) Variant monitoring has the ability to monitor in real-time the weekly trend in transmission of each variant in the world and in specific major countries, and evaluate the epidemic trend of these variants (Fig. 4B). This module calculates and analyzes the proportion of each variant at

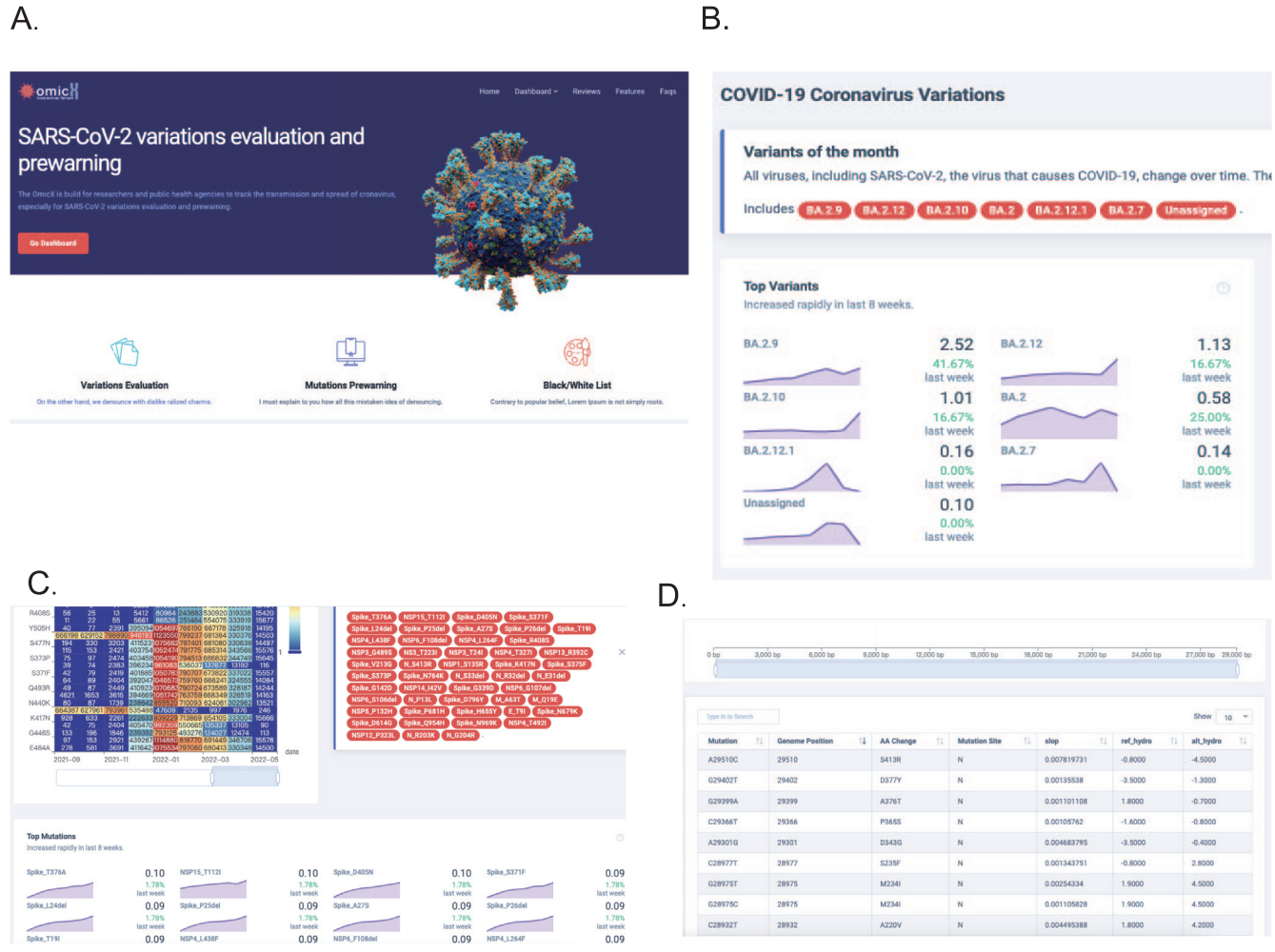


**Fig. 4.** Features of the mutation and variant monitoring and pre-warning system (MVMPS). A is the website portal. B is the variant monitoring module, which shows the global distribution of each variant by time frame and geography. C is the mutation monitoring module, which shows the global distribution of each mutation by time frame and geography. D is the mutation blacklist and whitelist module.

different times and different places. Thus, through analyzing changes in variant proportions, the possibility of a variant becoming a pandemic variant could be determined. For example, we found that in April 2022, the proportion of variant BA.2.12 increased rapidly at a weekly growth rate of 4 percentage points, which implied that this variant had the potential to become a major epidemic variant.

(2) Mutation monitoring can monitor in real-time the trend in transmission of each mutation in the world and in specific major countries, evaluate the impact of each mutation on transmission, determine the mutations that have a positive impact on virus transmission, and then provide early pre-warning (Fig. 4C). Compared with variant monitoring, mutation monitoring is more sensitive because mutations occur before the virus becomes a new variant. The emergence of each variant is a result of the accumulation of many mutations. Thus, through mutation monitoring, some potentially important mutations could be found before the emergence of a new variant. Mutation monitoring can also detect some important mutations distributed in different variants, which are usually produced by different variants through convergent evolution. For example, through mutation monitoring, we found that the proportion of Spike protein L452 in SARS-CoV-2 rapidly increased in April 2022. These results suggest that a mutation at L452 may help to improve the transmissibility of the virus or enable the virus to escape host immunity. The latest research by Xie et al. showed that mutations at L452 can enable the virus to escape host immunity.[32] Our further analysis found that L452 mutations in Spike protein mainly exist in four variants: L452M (BA. 2.13), L452R (BA.4 / BA.5), and L452Q (BA. 2.12). These results imply that these variants may obtain mutations at L452 through convergent evolution.

(3) The mutation blacklist and whitelist can be updated in real-time according to mutation monitoring and variant monitoring information (Fig. 4D). Generally, mutations affect gene function through their effect on protein structure. Therefore, it is of great value to evaluate the impact of each mutation on protein structure. In future work, we will add a structural model to the SARS-CoV-2 mutation and variant monitoring and pre-warning system (MVMPS), especially a structural model of Spike protein, and evaluate the impact of each mutation on protein structure. We believe that such a model will have important applications in the fields of gene function research, host adaptation, and drug discovery.

## 4. Conclusion

This study identified six base substitutions that reached saturation in the SARS-CoV-2 population. This means that currently undetected mutations will not be detected even if more genomes are sequenced. These undetected mutations were classified as the "mutation blacklist", and it is fair to assume that they have potential deleterious effects on the survival, replication, and transmission of SARS-CoV-2. Therefore, this mutation blacklist is of great value for the development of novel therapeutics to control and prevent SARS-CoV-2. The loci of mutations on the "mutation blacklist" likely play key roles in gene function, and further investigation may therefore deepen our understanding of these genes. Furthermore, the loci of mutations on the "mutation blacklist" are also conserved "weak spots" of SARS-CoV-2, which can be targeted by new drugs or vaccines. Analysis of the association between DNMs and transmission rate identified 185 DNMs that were positively correlated with virus transmission. These were classified as the "mutation whitelist". These mutations seem to benefit viral transmission and could be used to identify new variants with significant epidemic potential. To improve real-time monitoring of mutations and variants of SARS-CoV-2, and to update the mutation blacklist and whitelist, we built a SARS-CoV-2 mutation and variant monitoring website, which can evaluate the epidemic trend of mutations and variants and provide pre-warning for the prevention and control of SARS-CoV-2.

## CRediT authorship contribution statement

**Yamin Sun:** Writing – original draft, Visualization. **Min Wang:** Writing – original draft, Visualization. **Wenchao Lin:** Writing – original draft. **Wei Dong:** Software, Data curation. **Jianguo Xu:** Supervision, Writing – review & editing.

## Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This study was supported by funding from the Foundation of the Committee on Science and Technology of Tianjin (19YFZCSN00080), the State Key Research and Development Plan (2019YFC1605004), and the National Key Programs for Infectious Diseases of China (2017ZX10303405-001).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jobb.2022.06.006.

## References

1. Pal M, Berhanu G, Desalegn C, Kandi V. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): an update. *Cureus*. 2020;12(3).
2. Nakagawa S, Miyazawa T. Genome evolution of SARS-CoV-2 and its virological characteristics. *Inflamm Regener*. 2020;40(1):1–7.
3. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020.
4. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*. 2020;395(10224):565–574.
5. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265–269.
6. Zhou P, Yang X-L, Wang X-G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270–273.
7. WHO. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/.
8. Deng S, Xing K, He X. Mutation signatures inform the natural host of SARS-CoV-2. *Natl Sci Rev*. 2022;9(2):nwab220.
9. Harvey WT, Carabelli AM, Jackson B, et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol*. 2021;19(7):409–424.
10. Boehm E, Kronig I, Neher RA, Eckerle I, Vetter P, Kaiser L. Novel SARS-CoV-2 variants: the pandemics within the pandemic. *Clin Microbiol Infect*. 2021;27(8):1109–1117.
11. Wang Y, Chen R, Hu F, et al. Transmission, viral kinetics and clinical characteristics of the emergent SARS-CoV-2 Delta VOC in Guangzhou, China. *EClinicalMedicine*. 2021;40 101129.
12. Duffy S. Why are RNA virus mutation rates so damn high? *PLoS Biol*. 2018;16(8):e3000003.
13. Tria F, Pompei S, Loreto V. Dynamically correlated mutations drive human Influenza A evolution. *Sci Rep*. 2013;3(1):1–7.
14. Koelle K, Cobey S, Grenfell B, Pascual M. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*. 2006;314(5807):1898–1903.
15. Sandberg TE, Salazar MJ, Weng LL, Palsson BO, Feist AM. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. *Metab Eng*. 2019;56:1–16.
16. Liu Y, Liu J, Plante KS, et al. The N501Y spike substitution enhances SARS-CoV-2 transmission. *BioRxiv*. 2021.
17. Muth D, Corman VM, Roth H, et al. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Sci Rep*. 2018;8(1):1–11.

18. Ogando NS, Zevenhoven-Dobbe JC, van der Meer Y, Bredenbeek PJ, Posthuma EJ, Snijder EJ. The enzymatic activity of the nsp14 exoribonuclease is critical for replication of MERS-CoV and SARS-CoV-2. *J Virol*. 2020;94(23):e01246–e10320.
19. Sun Y, Lin W, Dong W, Xu J. Origin and evolutionary analysis of the SARS-CoV-2 Omicron variant. *J Biosaf Biosecur*. 2022;4(1):33–37.
20. Wang Y, Zhang Y, Chen J, et al. Detection of SARS-CoV-2 and its mutated variants via CRISPR-Cas13-based transcription amplification. *Anal Chem*. 2021;93(7):3393–3402. https://doi.org/10.1021/acs.analchem.0c04303.
21. Wang Y, Xue T, Wang M, et al. CRISPR-Cas13a cascade-based viral RNA assay for detecting SARS-CoV-2 and its mutations in clinical samples. Sens Actuators B Chem. 2022;362:131765. doi:10.1016/j.snb.2022.131765.
22. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*. 2017;22(13):30494.
23. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall*. 2017;1(1):33–46.
24. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24–26.
25. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80–92.
26. Sun Y, Wang M, Lin W, Dong W, Xu J. Massive-scale genomic analysis reveals SARS-CoV-2 mutations characteristics and evolutionary trends. mLife. In Print.
27. O'Toole Á, Scher E, Underwood A, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol*. 2021;7(2):veab064.
28. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261–272.
29. Yi K, Kim SY, Bleazard T, Kim T, Youk J, Ju YS. Mutational spectrum of SARS-CoV-2 during the global pandemic. *Exp Mol Med*. 2021;53(8):1229–1237.
30. Wang X, Zhao Y, Yan F, et al. Viral and host transcriptomes in SARS-CoV-2-infected human lung cells. *J Virol*. 2021;95(18):e00600–e621. https://doi.org/10.1128/JVI.00600-21.
31. Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G, Sanjana NE. The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. *Elife*. 2021;10:e65365.
32. Cao Y, Yisimayi A, Jian F, et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. bioRxiv. 2022:2022.04.30.489997. doi:10.1101/2022.04.30.489997.