**VIVID: An Integrated System to Closely Monitor and Predict Emerging and High-risk SARS-CoV-2 Variants**

Yupeng Li 1, Ying Huang 1, Jason Zhang 1

1 RVAC Medicines, Waltham, MA 02451, USA

## Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes continuously evolve, causing fresh waves of infections. There is a need to monitor new variants closely in order to predict whether these emerging variants will cause new pandemic waves. Multiple strategies have been implemented, including the collection of genome sequences, interactive dashboards monitoring the spread of variants, timely aggregated assessments from experts, and advanced machine learning algorithms to predict high-risk mutations and variants. We have developed a platform, called **VIVID** (Virus Variant Dissector), that integrates multiple resources to streamline the process of monitoring, predicting, and retrieving information on emerging variants, enabling quick responses to develop new or assess existing vaccines and therapeutics that can fight against these new variants. VIVID was developed using the R Shiny Dashboard and can be accessed at <http://vivid.rvacmed.com>. With the interactive platform, users can easily identify new emerging variants based on their own definition of emergence, closely follow the WHO weekly epidemiological updates, assess the new variants’ fitness based on an advanced prediction algorithm, and retrieve the spike gene sequences of the variants of interest for downstream experiments. In addition, the database is weekly updated, and the prediction algorithm is also retrained with the updated data to avoid performance drift. With integrated resources and tools, VIVID provides a centralized platform for researchers to closely monitor and quickly respond to the emerging SARS-CoV-2 variants.

**VIVID: 一个系统化监测和预测新冠变种流行的平台**

李玉鹏1, 黄颖1, 张敬新 1

1 RVAC Medicines, Waltham, MA 02451, USA

## 摘要

新型冠状病毒（SARS-CoV-2）基因组持续进化，不断产生新的、更具传播性的变种，从而造成疫情在全球多轮的蔓延。因此，及时侦测、研究、甚至提前预测新冠变种的流行可以帮助抑制新疫情，并使应对措施在时间上跑赢病毒进化。目前已有多个相关的工具和研究，包括新冠基因组的收集与分类、监测新冠变种流行的数据可视化平台、流行病学专家编纂的每周公告、以及预测高危突变和变种的机器学习算法。为了使研究工作者更好的利用这些工具，我们开发了一个名为VIVID(Virus Variant Dissector)的数据平台，系统化地整合了相关数据、工具和算法。该平台可以提供从监测到预测，再到序列提取等一体化功能，同时利用交互式数据可视化和简洁的用户界面，以提高易用性，从而使得研究者能够更方便、更主动地去应对新冠变种的流行，为下一步的疫苗和药物开发提供帮助。此外，后台数据和预测模型也会每周更新，以保持及时准确。

第一作者简介：

李玉鹏，2015年博士毕业于美国乔治亚大学，并在礼来、默沙东等制药公司从事多年生物信息和数据科学的研究和应用。2022年起入职RVAC Medicines，综合运用数据科学和人工智能进行mRNA疫苗的开发。

联系方式：[Yupeng.li@rvacmed.com](mailto:Yupeng.li@rvacmed.com)

## Introduction

The ongoing pandemic of coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has posed overwhelming challenges to public health, virology research, and drug discovery and development. The high mutation rate exhibited by the SARS‑CoV‑2 genome allows for the emergence of new variants that are more transmissible and more capable of evading the host's immune responses. Multiple variants of concern (VOCs) have caused resurgent outbreaks of COVID-19 worldwide: Alpha and Beta in December 2020, Gamma in January 2021, Delta in May 2021, and Omicron in November 2021. Variant surveillance is urgently needed to understand the genome evolution in order to respond to new high-risk variants quickly. Thanks to the great contributions and collaborations from the scientific and public healthcare communities, many resources have become available for variant monitoring. One of the most important resources is the Global Initiative on Sharing Avian Influenza Data (GISAID), a virus data-sharing initiative to help researchers understand how viruses evolve and spread during pandemics 1,2. To date, over 12 million SARS-CoV-2 genomic sequences from 194 countries and territories have been submitted to GISAID’s EpiCoV database. CoV-Lineages, part of the Pango dynamic nomenclature system for identifying SARS-CoV-2 genetic lineages of epidemiological relevance, provides a rational and dynamic virus nomenclature to further classify those genomes into Pango lineages3. Various visualization tools utilizing the GISAID genomic sequences and Pango lineage assignments have greatly helped broader communities to easily monitor and understand the SARS-CoV-2 variants, e.g., Nextstrain 4, outbreak.info 5, and Cov-SPECTRUM 6.

In addition to monitoring for the emergence of new variants closely, many efforts have been made to predict which variants or mutations will spread widely and cause new waves of SARS-CoV-2 infection so that healthcare authorities and researchers can proactively prevent or respond to these new outbreaks 7–12. Three types of features are used for predicting the mutation risk: epidemiology, evolution, and immunity. Epidemiology features include the observed variant frequencies and spreading patterns in the human population. Time-series analysis can forecast the future spreading trend from the epidemiology features and does not rely on any assumptions of underlying molecular mechanisms. Evolutionary features, calculated from a phylogenetic tree of existing protein or genomic sequences, measure the selection pressure at each nucleotide or amino acid site, with the hypothesis that mutations in regions under selective pressure from the immune system could escape the immune response and cause a new spread of infections. Deep-learning language models have been used to understand the amino acid embedding patterns, which estimate the probability of a mutation given the surrounding sequence context and thus can be considered as evolution features 7,9. Immunity features measure the effect of mutations on the immune response. For example, deep mutational scanning can experimentally detect the mutation effect on binding to the angiotensin-converting enzyme 2 (ACE2) receptor and antibody escape from a large mutagenesis library. If a new variant has many mutations with stronger ACE2 binding affinity or antibody escape capability, the variant will potentially spread widely with increased transmission and immune evasion 13,14.

All these studies and resources provide unprecedented opportunities to identify and predict emerging SARS-CoV-2 variants to enable scientists to quickly follow up and develop more preventative vaccines or effective therapeutics against COVID-19. However, it is not easy for scientists to compile, select, and utilize all relevant resources. In addition, the prediction algorithm using historical data cannot accurately predict new variants due to their nearly constant mutational evolution. Additionally, quick and accurate retrieval of the genetic sequences for study after identification the variant(s) of interest remains challenging and burdensome. Therefore, we have integrated multiple useful resources and developed VIVID, a centralized platform that enables scientists to closely monitor the prevalence of variants, predict high-risk variants with weekly updated data, and easily retrieve spike gene sequences. We have also made it easy to use by utilizing interactive dashboards and a simplified user interface (UI).

## Results

**Epidemiological monitoring**

The platform consists of three major components: epidemiological monitoring, fitness prediction, and spike gene sequence retrieval. Epidemiological monitoring mainly focuses on the VOCs designated by public health institutes, e.g., the World Health Organization (WHO) and the United States Centers for Disease Control and Prevention (US CDC). The data come from outbreak.info 5, which timely processes the raw genomic sequences from GISAID and summarized the prevalence of VOCs, Pango lineages 3, and mutations. VIVID retrieves the global prevalence data on a weekly basis via the application programming interface (API) provided by outbreak.info and presents the information in multiple R Shiny dashboards. Users can visualize the trend lines and mutation frequency of prevalent VOCs and their Pango lineages and interactively zoom into specific trend lines and time ranges. Moreover, users can select the prevalence threshold to define and prioritize the recently prevalent VOCs and lineages (Figure 1).

a

b

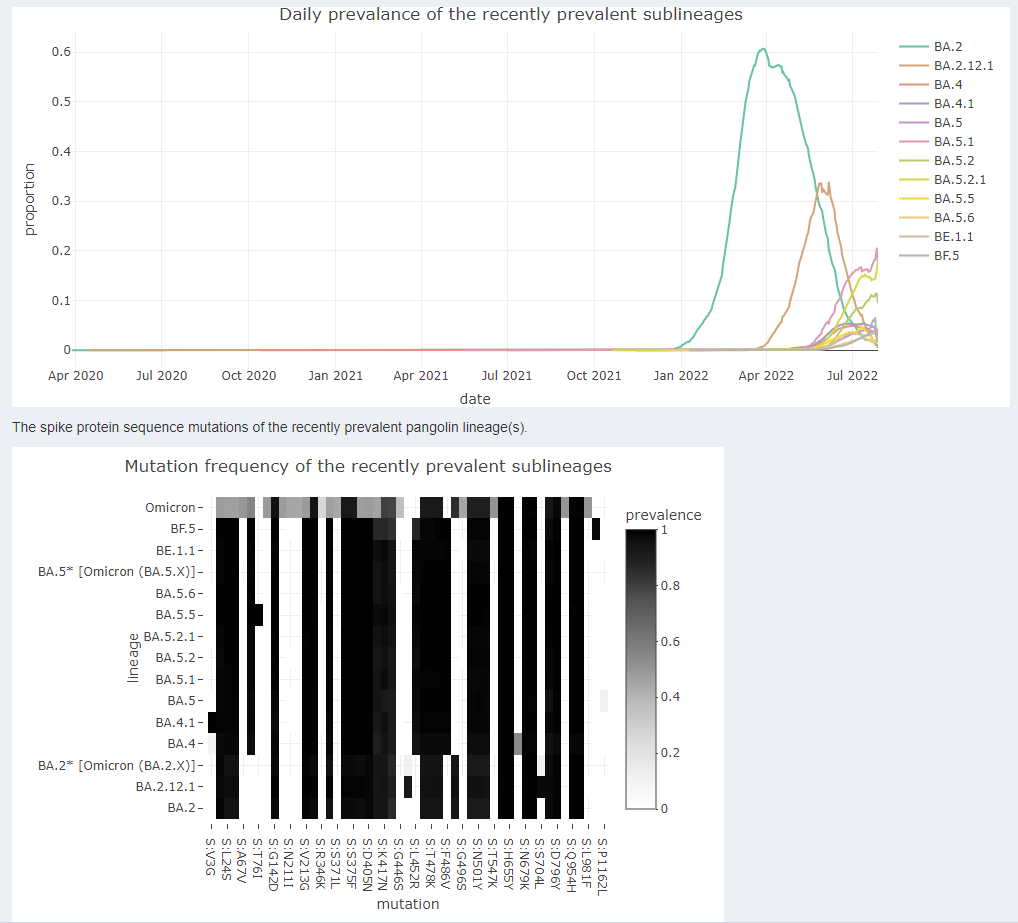
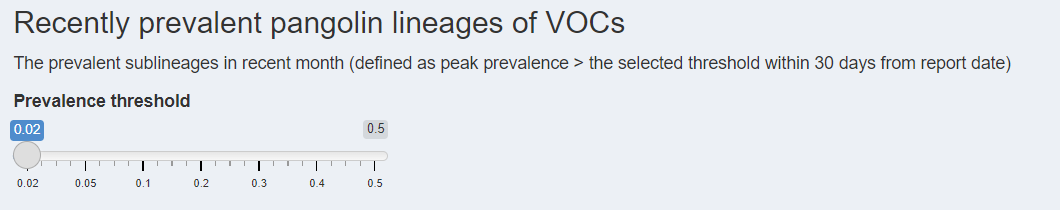
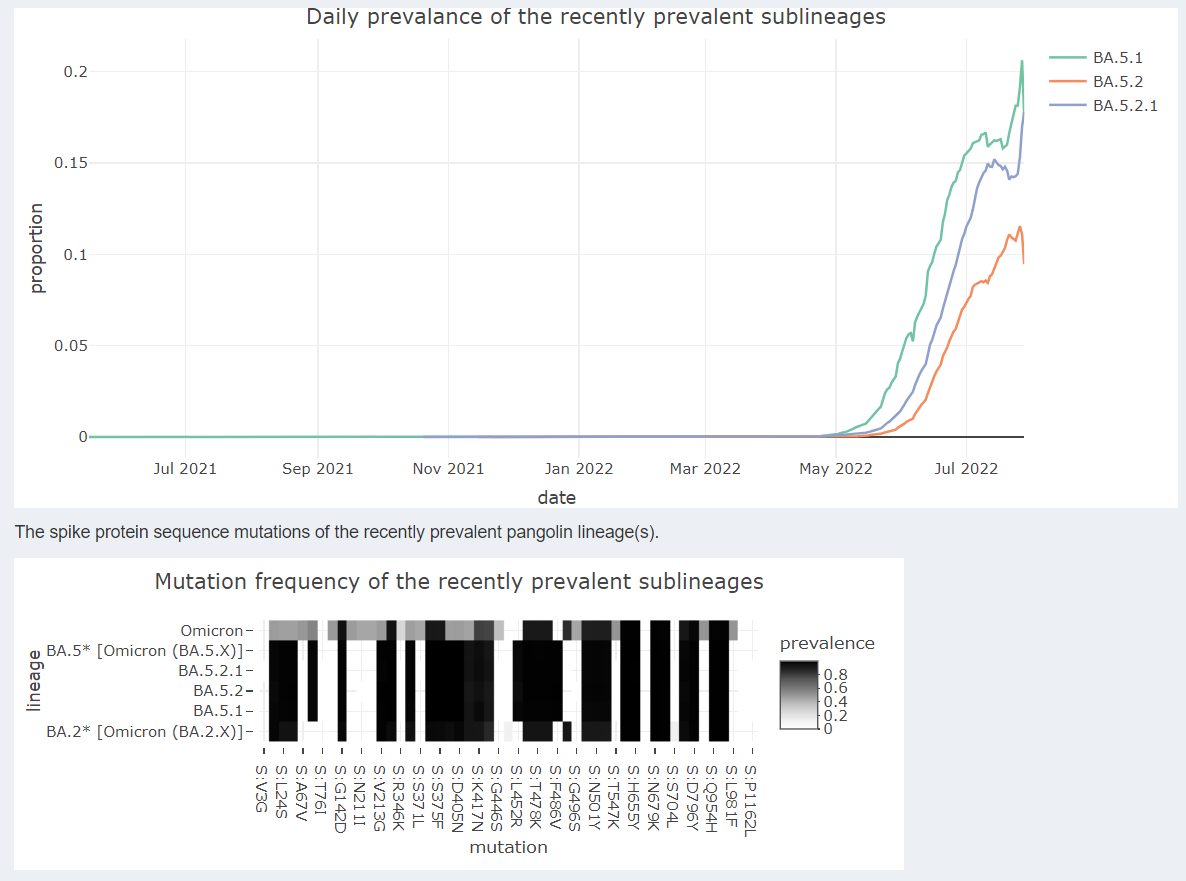
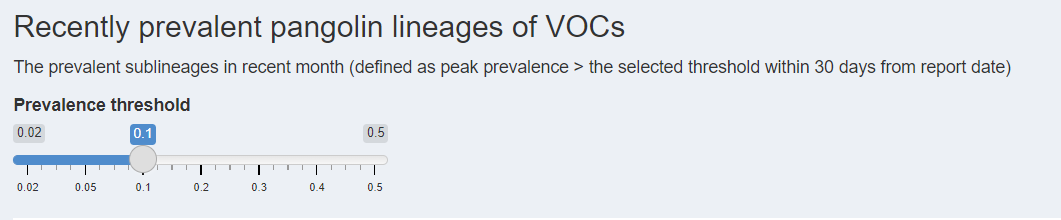


Figure 1. The R Shiny dashboards of lineage prevalence and mutation frequency. Notably, the lineages will automatically change depending on the prevalence threshold. a) Prevalence threshold = 0.02. b) Prevalence threshold = 0.1.

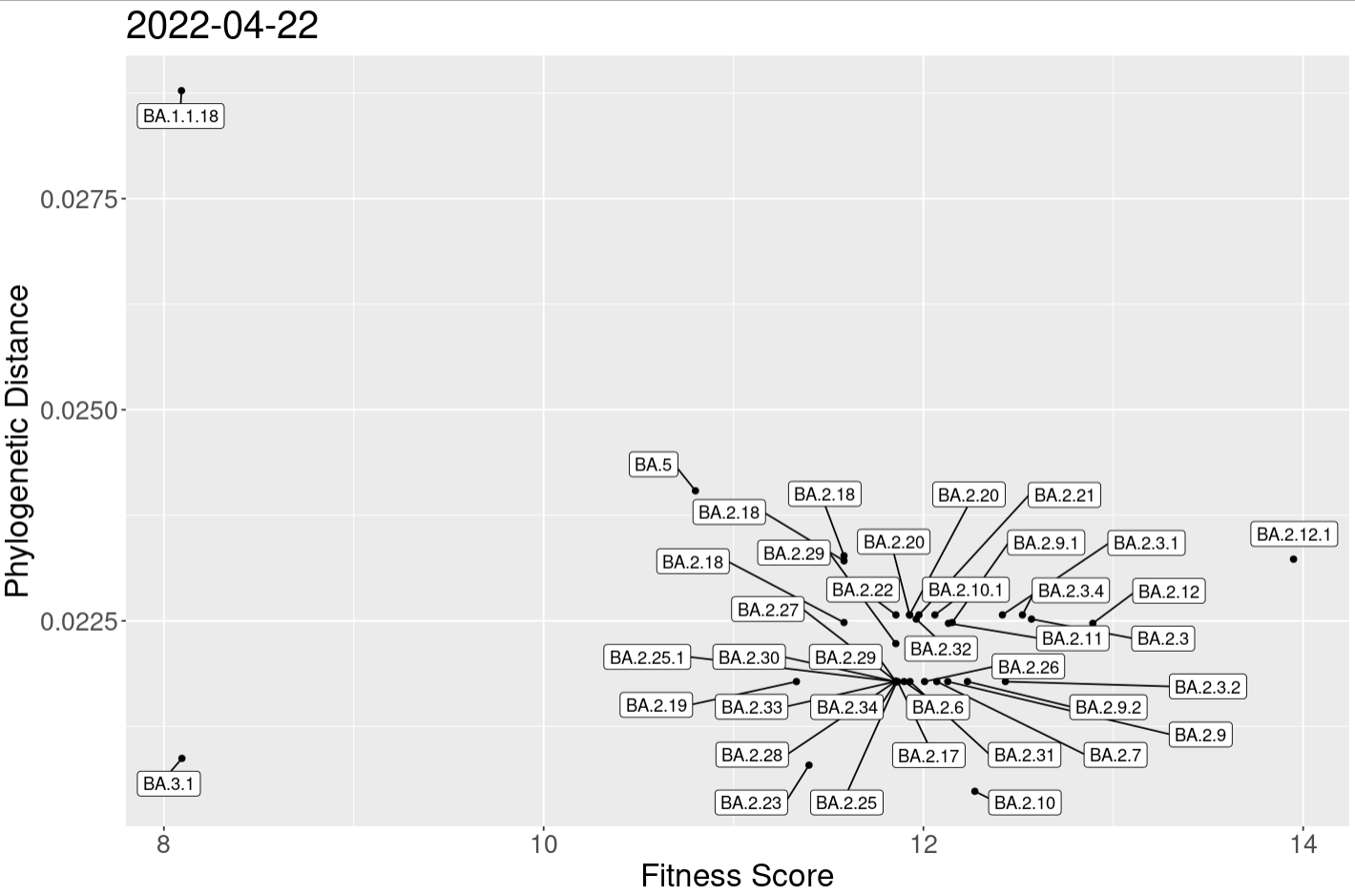
**Fitness prediction**

Among the three types of features used for predicting high-risk variants or mutations, it was shown that epidemiology features are the most robust and effective 10. We implemented a powerful and flexible algorithm to predict the fitness of both variants and mutations 11 by fitting a Bayesian linear regression to the prevalence of lineages. The algorithm estimates each lineage’s relative fitness (basic reproduction number, R0) and then calculates the contribution of individual mutations to the lineage fitness. Instead of using handcrafted features to measure lineage fitness or emergence, it fits a lineage’s prevalence trend over time as a logistic growth function. The growth rate in the estimated logistic growth function can be considered as the fitness score. With the estimated parameters of the growth function, we can forecast the future prevalence of the lineage. Furthermore, the growth rate was modeled as a linear combination of the effects of individual mutations, i.e., individual mutation fitness scores, so mutation fitness can be retrieved from the linear model and further be used to predict the overall fitness of an unseen lineage or new lineage without much sequence data. The prediction is generally accurate and robust, and able to reliably forecast 1-2 months into the future for VOCs.

As Obermeyer et al 11 demonstrated, and which we also observed in all variant prediction algorithms, a completely new lineage with multiple novel mutations, e.g., Omicron BA.1 and BA.2, can disrupt the trends of existing lineages, causing inaccuracies in their predictive nature. Nevertheless, the algorithm we implemented can be updated regularly with new data and become more stable and accurate within two weeks after the appearance of the disrupting lineages. Therefore, we update VIVID with new genomic data on a weekly basis and recalibrate the algorithm to enable predictions based on the most up‑to‑date real-world data. The genomic data come from public SARS-CoV-2 genome sequences aggregated from GenBank, COG-UK, and the China National Center for Bioinformation 15. The VIVID platform lists the ranked fitness scores for both lineages and mutations from the training results every week.

When the algorithm was published in May 2022, the authors correctly inferred Omicron variants (BA.1 and BA.2 lineages) to have significantly higher fitness than others using sequence data up to February 2022 11, but didn’t have the chance to update their results with the new sequence data when new variants emerged after publication With the same algorithm but weekly updated sequence data, we further detected new high-risk lineages, BA.2.12.1, BA.4, and BA.5, about 2 to 3 weeks before they became globally dominant lineages (Figure 2).

a



b

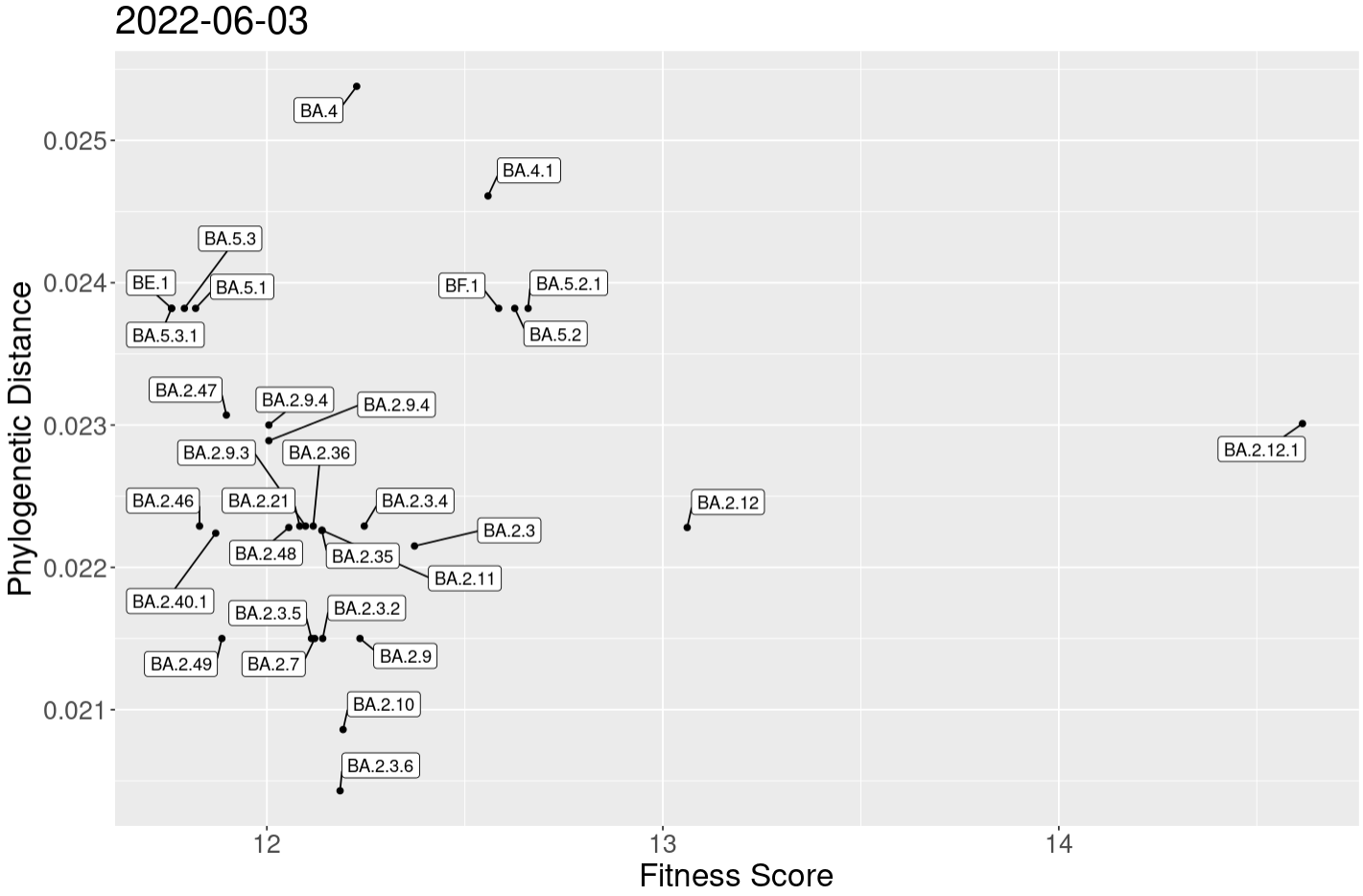


Figure 2. The relative distance and estimated fitness of top fitted lineages at different time points. The relative fitness score is the fold increase in fitness over the reference lineage (A lineage). The relative distance is the lineage’s phylogenetic distance to the reference lineage. The phylogenetic trees were built only using spike protein sequences. a) BA.2.12.1 lineage, 97% of its sequences found in the US, first showed significantly higher fitness scores than other top fitted lineages in the weekly retrained results on 2022-04-22, two weeks before it became the dominant lineage in the US. b) BA.4\* and BA.5\* lineages first showed significantly higher fitness scores than other top fitted lineages on 2022-06-03, two weeks before they became the globally dominant lineages.

**Spike gene sequence retrieval**

After examining the monitoring and prediction results, users may identify a lineage of interest, e.g., BA.5, and want to obtain its spike gene sequence for follow-up investigation. To our knowledge, there is no existing tool that allows users to easily retrieve SARS-CoV-2 spike gene sequences based on lineages. In addition, thousands of genomic sequences can be assigned to the same lineage, so we will need to select one genomic sequence or generate a consensus spike gene sequence for the lineage of interest, which is not an easy task, especially for bench scientists. Depending on the purpose of the follow-up study, there are different ways to select for the sequence of interest. For example, users may select the oldest, newest, or most frequent sequences of the lineage. Users may also choose to include all descendant lineages and sequences of the specified lineage. Users may be interested in specific sequence types, e.g., mRNA, cDNA, and protein. The platform provides a flexible and convenient tool to retrieve the consensus spike gene sequences. Users can specify the lineages, the method to select the sequence, and sequence types, and then the identified sequences with FASTA format will be displayed immediately (Figure 3). Additionally, the spike gene sequences, including cDNA and protein, are downloaded from GISAID weekly to keep the most up-to-date sequences, and the frequency and collection date of unique sequences for each lineage are pre-calculated so the platform can respond to users’ queries instantly.

In addition to sequence retrieval from specified lineages, the platform also allows users to identify the lineage by providing a spike gene sequence. Figure 3 displays a sample Basic Local Alignment Search Tool (BLAST) search result with top 5 hits for users to identify the potential lineage.

Graphical user interface, text, application

Description automatically generated

Figure 3. The user interface of spike gene sequence retrieval from specified Pango lineages

## Discussion

Timely monitoring and predicting emerging virus variants are critical for developing new vaccines and therapeutics that can fight against the SARS-CoV-2 variants. VIVID, the integrated platform we have developed will greatly improve the process by providing weekly updated prevalence visualizations, fitness prediction, and gene sequence retrieval, all of which enable improved efficiencies and more proactive responses to the threat of additional waves of COVID-19 outbreaks due to emerging SARS‑CoV-2 variants.

We will continue to improve the functionality and UI of the VIVID platform. For instance, the addition of geographic information will allow users to monitor high-risk lineages before they become highly prevalent. Inclusion of social network posts and news articles will also provide good resources to monitor high-risk lineages before they become highly prevalent. The world first became aware of BA.2.75, a lineage with multiple mutations that have antibody escaping potential, through social network posts early on when there were only a few dozen BA.2.75 genome sequences available. For the fitness prediction, assessments using genomic sequences from National Center for Biotechnology Information (NCBI), dominated by sequences from the United Kingdom and the United States, may miss or delay the prediction when a lineage emerges from other countries first, e.g., South Africa and India. GISAID has more diverse data, but the lineage assignment is often delayed or incorrect. Creating an optimal balance between data diversity and prediction accuracy is worth further investigation. The prediction algorithm is based on linear regression, which may miss the interaction effect between mutations. The algorithm can be improved using tree- or artificial neural network-based methods. Furthermore, the VIVID platform provides multiple tools and resources for users to make decisions by themselves. However, it may be more convenient if VIVID can automatically detect new emerging and high-risk lineages, summarize weekly results, and alert the audience. Finally, the VIVID platform is not limited to SARS-CoV-2. Theoretically, we can expand VIVID to include data and information for other viruses, e.g., influenza and monkeypox, thereby providing a powerful tool for healthcare authorities and researchers to combat ongoing and future epidemics and pandemics.

## References

1. Khare, S. *et al.* GISAID’s Role in Pandemic Response. *China CDC Weekly* **3**, 1049 (2021).

2. Elbe, S. & Buckland‐Merrett, G. Data, disease and diplomacy: GISAID’s innovative contribution to global health. *Global Challenges* **1**, 33 (2017).

3. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology 2020 5:11* **5**, 1403–1407 (2020).

4. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

5. Tsueng, G. *et al.* Outbreak.info Research Library: A standardized, searchable platform to discover and explore COVID-19 resources. *bioRxiv* 2022.01.20.477133 (2022) doi:10.1101/2022.01.20.477133.

6. Chen, C. *et al.* CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics* **38**, 1735–1737 (2022).

7. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).

8. Pucci, F. & Rooman, M. Prediction and Evolution of the Molecular Fitness of SARS-CoV-2 Variants: Introducing SpikePro. *Viruses 2021, Vol. 13, Page 935* **13**, 935 (2021).

9. Beguir, K. *et al.* Early Computational Detection of Potential High Risk SARS-CoV-2 Variants. *bioRxiv* 2021.12.24.474095 (2021) doi:10.1101/2021.12.24.474095.

10. Maher, M. C. *et al.* Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Science Translational Medicine* **14**, 3445 (2022).

11. Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).

12. Pucci, F. & Rooman, M. Prediction and Evolution of the Molecular Fitness of SARS-CoV-2 Variants: Introducing SpikePro. *Viruses 2021, Vol. 13, Page 935* **13**, 935 (2021).

13. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science*  **371**, 850–854 (2021).

14. Starr, T. N. *et al.* Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *BioRxiv* (2022).

15. Turakhia, Y. *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics 2021 53:6* **53**, 809–816 (2021).