

Quantifying the immunological distinctiveness of emerging SARS-CoV-2 variants in the context of prior regional herd exposure

Michiel J. M. Niesen¹, Karthik Murugadoss¹, Patrick J. Lenehan¹, Aron Marchler-Bauer², Jiyao Wang², Ryan Connor², J. Rodney Brister², A. J. Venkatakrishnan¹ and Venky Soundararajan¹*

¹nference, One Main St, East Arcade, Cambridge, Massachusetts 02139, USA

²National Center for Biotechnology Information, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

*To whom correspondence should be addressed: Email: venky@nference.net

Edited By: Karen E. Nelson

Abstract

The COVID-19 pandemic has seen the persistent emergence of immune-evasive SARS-CoV-2 variants under the selection pressure of natural and vaccination-acquired immunity. However, it is currently challenging to quantify how immunologically distinct a new variant is compared to all the prior variants to which a population has been exposed. Here, we define “Distinctiveness” of SARS-CoV-2 sequences based on a proteome-wide comparison with all prior sequences from the same geographical region. We observe a correlation between Distinctiveness relative to contemporary sequences and future change in prevalence of a newly circulating lineage (Pearson $r = 0.75$), suggesting that the Distinctiveness of emergent SARS-CoV-2 lineages is associated with their epidemiological fitness. We further show that the average Distinctiveness of sequences belonging to a lineage, relative to the Distinctiveness of other sequences that occur at the same place and time ($n = 944$ location/time data points), is predictive of future increases in prevalence (Area Under the Curve, AUC = 0.88 [95% confidence interval 0.86 to 0.90]). By assessing the Delta variant in India versus Brazil, we show that the same lineage can have different Distinctiveness-contributing positions in different geographical regions depending on the other variants that previously circulated in those regions. Finally, we find that positions that constitute epitopes contribute disproportionately (20-fold higher than the average position) to Distinctiveness. Overall, this study suggests that real-time assessment of new SARS-CoV-2 variants in the context of prior regional herd exposure via Distinctiveness can augment genomic surveillance efforts.

Significance Statement:

Variants of SARS-CoV-2 have continually emerged under the selection pressure of host immunity. Though there are various phylogenetic techniques to study SARS-CoV-2 diversity, to our knowledge, there is no metric that captures how immunologically distinct a new SARS-CoV-2 variant is compared to the prior variants seen in a geographical region. The Distinctiveness metric that we have developed gives an intuitive understanding of the immunological distinctiveness of a given variant and relative Distinctiveness is strongly correlated to competitive fitness. The use of Distinctiveness in real-time assessment of viral genomes holds promise for aiding pandemic preparedness initiatives.

Introduction

To date, over 10 billion COVID-19 vaccine doses have been administered globally (1), with over 200 million individuals fully vaccinated in the United States (2). Recent studies have confirmed that natural immunity (i.e. immunity gained through prior infection) is also highly protective and may even provide more durable protection than vaccination alone (3–12). Given that over 400 million COVID-19 cases have been reported worldwide (with over 78 million cases in the United States) (1), it is likely that both vaccination-acquired immunity and natural immunity play important roles in the evolution of new SARS-CoV-2 variants.

Throughout the course of the COVID-19 pandemic, SARS-CoV-2 has evolved to generate new variants which harbor unique constellations of mutations (substitutions, deletions, and insertions). Some of these variants are designated as Variants of Concern (VOCs) based on evidence for increased transmissibility, increased disease severity, or reduced neutralization by vaccine-elicited sera or authorized monoclonal antibody treatments. Such variants include Alpha (B.1.1.7 and Q lineages per PANGO classification), Beta (B.1.351 and descendants), Gamma (P.1 and descendants), Delta (B.1.617.2 and AY lineages), and most recently Omicron (B.1.1.529 and BA lineages) (13). As new SARS-CoV-2 variants evolve, it is im-

OXFORD
UNIVERSITY PRESS

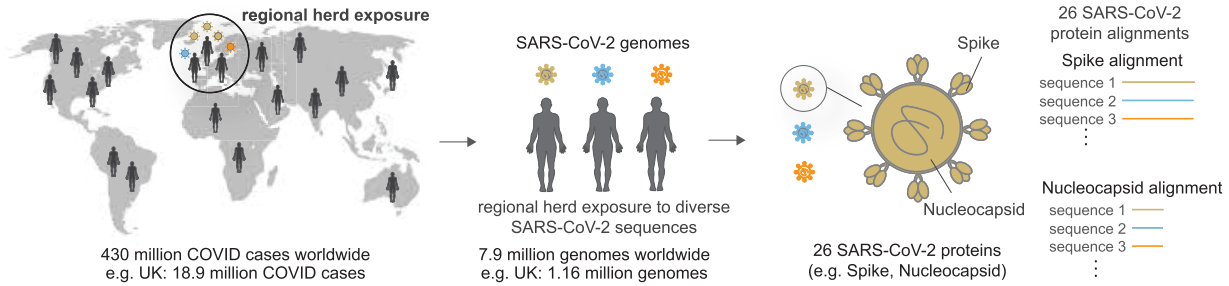
Competing Interests: M.N., K.M., A.V., P.L., and V.S. are employees of nference and have financial interests in the company. nference is collaborating with bio-pharmaceutical, medical device and diagnostics companies, public health agencies, academic medical centers, and health systems on data science initiatives unrelated to this study. These collaborations had no role in study design, data collection and analysis, decision to publish, or preparation of this manuscript. A.M.B., J.W., R.C., and J.B. declare no conflicts of interest.

[†]M.J.M.N., K.M., and P.J.L. contributed equally to this work.

Received: May 31, 2022. **Accepted:** June 29, 2022

© The Author(s) 2022. Published by Oxford University Press on behalf of National Academy of Sciences. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

a. Generation of geographic region-based amino acid sequence alignments of all 26 SARS-CoV-2 proteins to capture regional herd exposure



b. Comparison of mutational load and 'Distinctiveness' for a given SARS-CoV-2 sequence

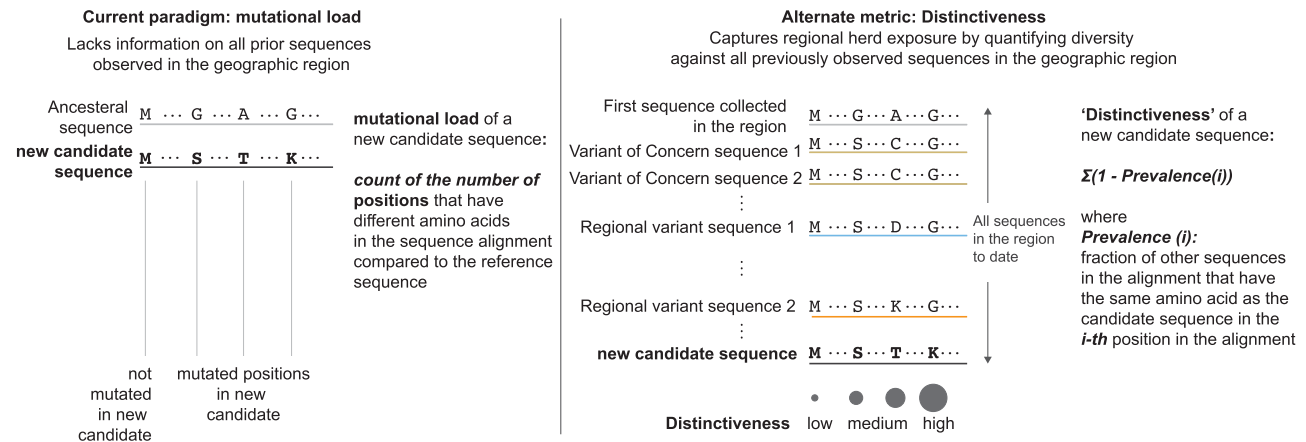


Fig. 1. (a) Generation of geographic region-based amino acid sequence alignments of all 26 SARS-CoV-2 proteins to capture regional herd exposure. (b) Comparison of mutational load and "Distinctiveness" for a given SARS-CoV-2 sequence.

portant to estimate their likelihood of evading existing regional herd immunity and potentially transmitting highly at the community level. While the main evidence that a variant evades immune response typically relies on laboratory assays and epidemiological evidence (14–20), complementary approaches that provide initial evidence, as soon as new sequences are reported, could enable earlier response.

Here, we define a new metric "Distinctiveness" to capture the proteome-level novelty of emerging SARS-CoV-2 sequences against all the documented regional lineages. Distinctiveness aims to quantify previous herd exposure to viral sequences that are similar to the current sequence, capturing an important factor of viral epidemiological fitness. This approach views viral evolution through a new lens that considers the pressure to evolve new strains harboring protein content to which communities have not previously been exposed. We show that the same lineage can have different Distinctiveness values simultaneously in different countries, as well as different Distinctiveness-contributing positions. We find that the relative Distinctiveness of emergent SARS-CoV-2 lineages is associated with their epidemiological fitness, as defined by the change in the lineage prevalence. We also show that epitope positions contribute disproportionately to Distinctiveness.

Results

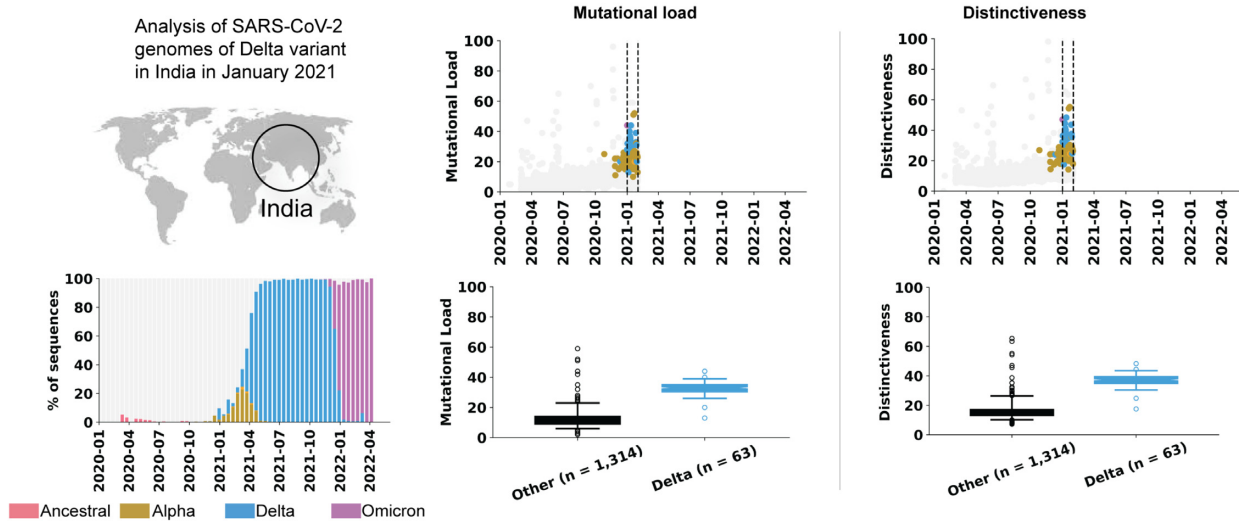
"Distinctiveness" as a metric to capture novelty of emerging SARS-CoV-2 sequences

Understanding the immunological novelty of a SARS-CoV-2 strain for a given population needs to take into consideration which sequences were previously seen at a regional level and for which

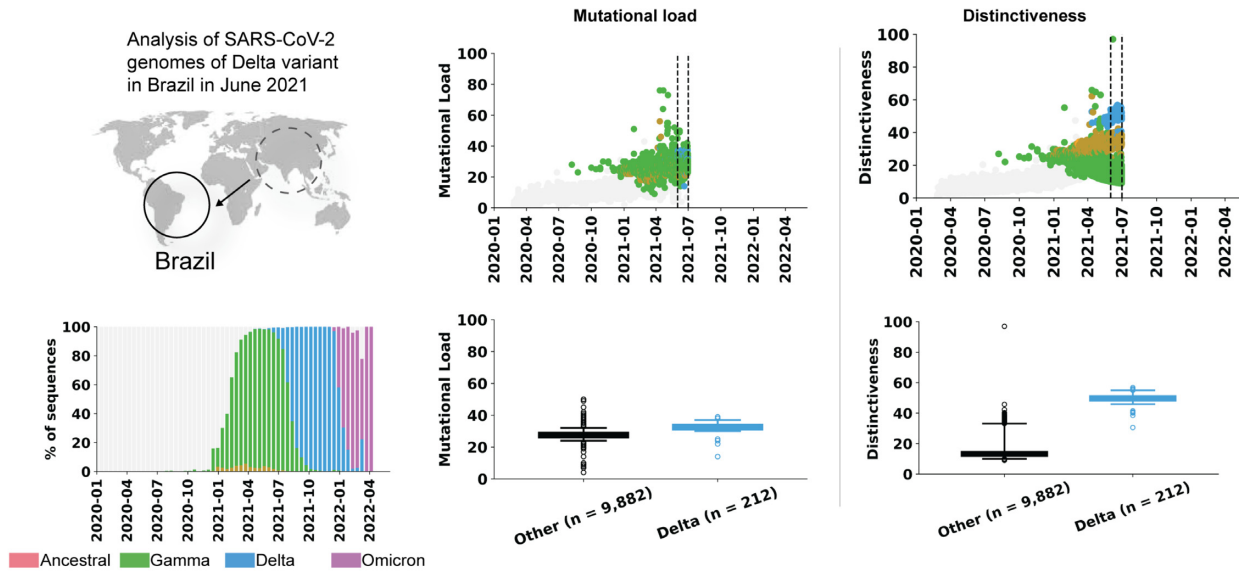
there might exist population-level immunity. Here, we introduce a new metric "Distinctiveness" of a given SARS-CoV-2 sequence based on comparison against all available sequences previously collected from the same region. Specifically, Distinctiveness is defined as the average distance at the amino-acid level between a sequence and all prior sequences (Figure 1; see the "Methods" section). Distinctiveness can be computed at the global level or at a regional level for any chosen time period. Below, we compare Distinctiveness of the VOCs with contemporary sequences and investigate the relationship between Distinctiveness of a sequence and the change in its regional prevalence. For comparison, we also report the "Mutational load" of the same sequences. Mutational load is simply defined as the number of mutations in the new sequence compared with the ancestral reference sequence (GenBank: MN908947.3), and as such it does not account for the entirety of SARS-CoV-2 evolution or the local prevalence of sequences.

We computed mutational load and Distinctiveness during the emergence of the VOCs in the country of their emergence. Both mutational load and Distinctiveness values of VOC sequences were significantly higher than contemporary lineages (Supplementary Figures S1 and 2). For example, we consider the emergence of the Delta variant in India during January 2021. Both mutational load and Distinctiveness of the Delta variant in India were significantly higher than that of the other contemporary lineages (Figure 2a). This raises the question of whether Delta variant sequences were also competitive in other countries. We considered the example of Brazil, where the Gamma variant was dominant prior to the arrival of the Delta variant (Figure 2b). Whereas the mutational load of the Delta variant was comparable to those of contemporary lineages, the Distinctiveness of the Delta variant

a. Comparison of mutational load and Distinctiveness during the emergence of the Delta variant in India



b. Comparison of mutational load and Distinctiveness during the first appearance of the Delta variant in Brazil



c. Positions that contribute to Distinctiveness differ by geographical region; due to differences in previously circulating sequences

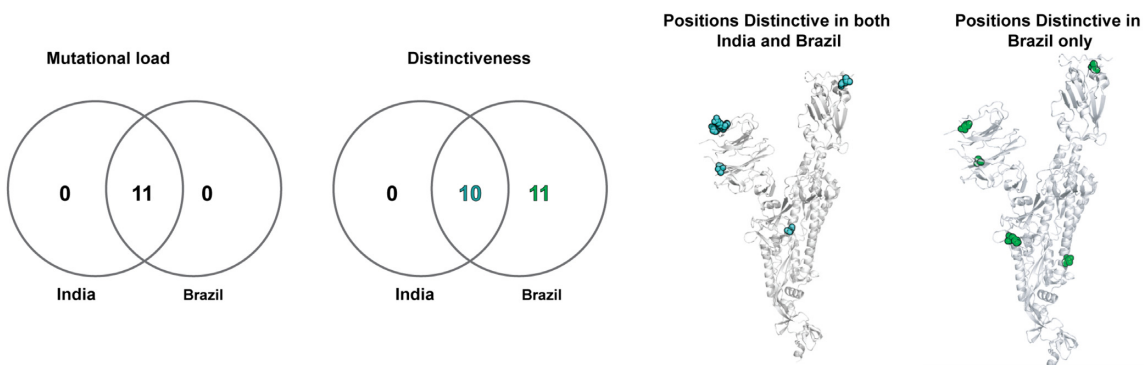
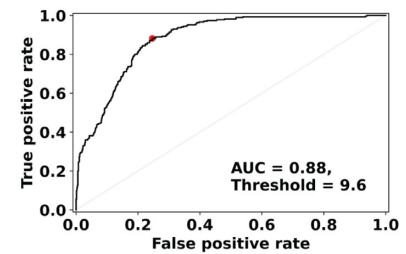
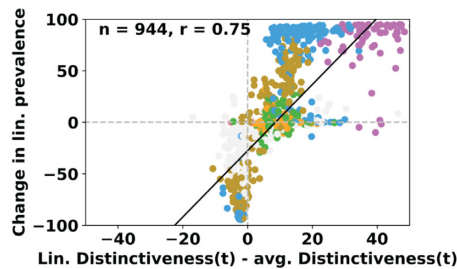


Fig. 2. Sequence Distinctiveness as a function of time in geographical regions where VOCs first emerged. Comparison of mutational load and Distinctiveness during the emergence of the Delta variant in (a) India and (b) Brazil. (c) Comparison of mutational load and Distinctiveness in Spike protein of Delta variant in India and Brazil. Venn diagrams compare the positions that contribute to Mutational load or high Distinctiveness (from Supplementary Figure S3). The positions that have high Distinctiveness are highlighted on the protein structures of the Spike protein as spheres (PDB identifier: 6VSB). (a)–(b) Sequences classified as VOCs are brightly colored dots (Alpha: brown, Beta: orange, Gamma: green, Delta: blue, Omicron: magenta) and other sequences are gray dots. Shown on the right is a comparison of the Distinctiveness values for the emerging VOC sequences and contemporary sequences, collected during the indicated time periods (inset and vertical dashed lines).

a. Correlation between the local Distinctiveness of a lineage and the future change in prevalence of that lineage

Analysis of 944 combinations of geographical regions and time-windows



b. Analysis of Distinctiveness using only protein positions associated with antibody binding

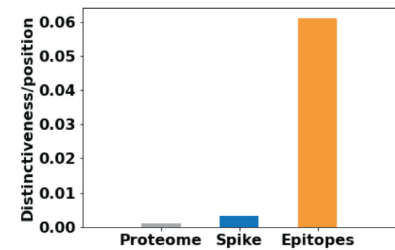
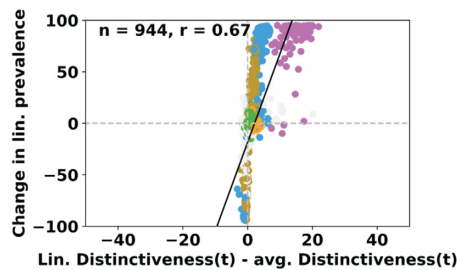
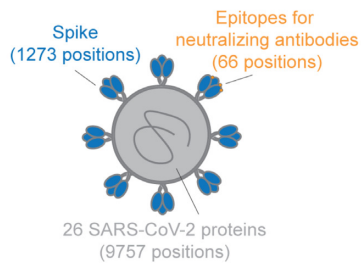


Fig. 3. The average sequence distinctiveness as a predictor of future changes in prevalence of a lineage. (a) Comparing the correlations of Distinctiveness of a lineage with its competitiveness in the geographic region (across 294 country/time data points). Distinctiveness of a lineage, relative to the average of all sequences that were collected from the same region during the same time, is predictive of future changes in prevalence. The ROC is shown for predicting an increase in prevalence of greater than 20 percentage points from an initial 28-d time window and a subsequent 28-d time window, starting 56 d in the future. (b) Distinctiveness calculated for only a subset of 66 positions involved in neutralizing antibody binding (orange) retains most of the predictive capacity. These positions were found to contribute disproportionately to the overall Distinctiveness, with ~20-fold higher average Distinctiveness as compared with average positions.

was significantly higher. Indeed, the Delta variant outcompeted the Gamma variant to become the dominant strain in Brazil (Figure 2b).

We next assessed which specific positions contribute most to the observed Distinctiveness values of the Delta variant in India and Brazil. We compared the mutational frequency and average Distinctiveness contribution for each amino acid position in the Spike protein of Delta variant sequences collected in India versus Brazil (Figure 2c and Supplementary Figure S3). In India, where the Delta variant originated, the 11 mutated positions correspond almost exactly to the Distinctiveness-contributing positions. The only exception is the 614 position on the Spike protein. This position has not contributed to the Delta variant's Distinctiveness as it has been highly prevalent globally (i.e. present in over 99% of SARS-CoV-2 genomes deposited in GISAID) since June 2020 (15, 21, 22). Brazil, on the other hand, experienced a large wave of cases dominated by the Gamma variant before the arrival of the Delta variant. Here, in addition to the same 10 Spike protein mutations that were observed in India (Supplementary Figure S3a), there were 11 other positions that further contributed to its regional Distinctiveness (Supplementary Figure S3b). These additional positions correspond to known Gamma lineage-defining mutations (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I, and V1176F).

Relative Distinctiveness of emergent SARS-CoV-2 lineages is associated with their epidemiological fitness

In order to examine a possible relationship between Distinctiveness and epidemiological fitness of SARS-CoV-2 lineages, we assessed the correlation between Distinctiveness and change in

prevalence for all circulating lineages (grouped as the VOCs and a single group combining all non-VOCs) in 78 geographical regions (27 countries and 51 US states). We find that the relative Distinctiveness of emergent SARS-CoV-2 lineages is associated with their change in lineage prevalence over eight weeks (Figure 3a and Supplementary Figure S4a) (Pearson $r = 0.75$). In comparison, mutational load was found to have a lower association with change in prevalence (Pearson $r = 0.53$). We further find that the average Distinctiveness of a lineage in a country/time window can predict future increases in prevalence (Figure 3a, AUC = 0.88 [95% confidence intervals (CIs) 0.86 to 0.90], for predicting a greater than 20 percentage point increase in local prevalence; Supplementary Figure S3b).

Positions that constitute epitopes contribute disproportionately to Distinctiveness compared to the overall SARS-CoV-2 proteome

Since Distinctiveness is intended to capture the fitness of a sequence in the context of previous herd exposure to similar sequences, we next investigated Distinctiveness in the context of known immunogenic positions. Specifically, we analyzed the Distinctiveness of only Spike protein positions and for 66 epitope positions, previously associated with neutralizing antibody binding or therapeutic agent binding (Figure 3b). We found that Distinctiveness determined using only the 66 epitope positions was still correlated with future changes in lineage prevalence (Pearson $r = 0.67$). Additionally, we found that both Spike positions (average Distinctiveness of 0.007/position) and the 66 epitope positions (average Distinctiveness of 0.061/position) exhibit increased contributions to overall Distinctiveness (average Distinctiveness of 0.003/position).

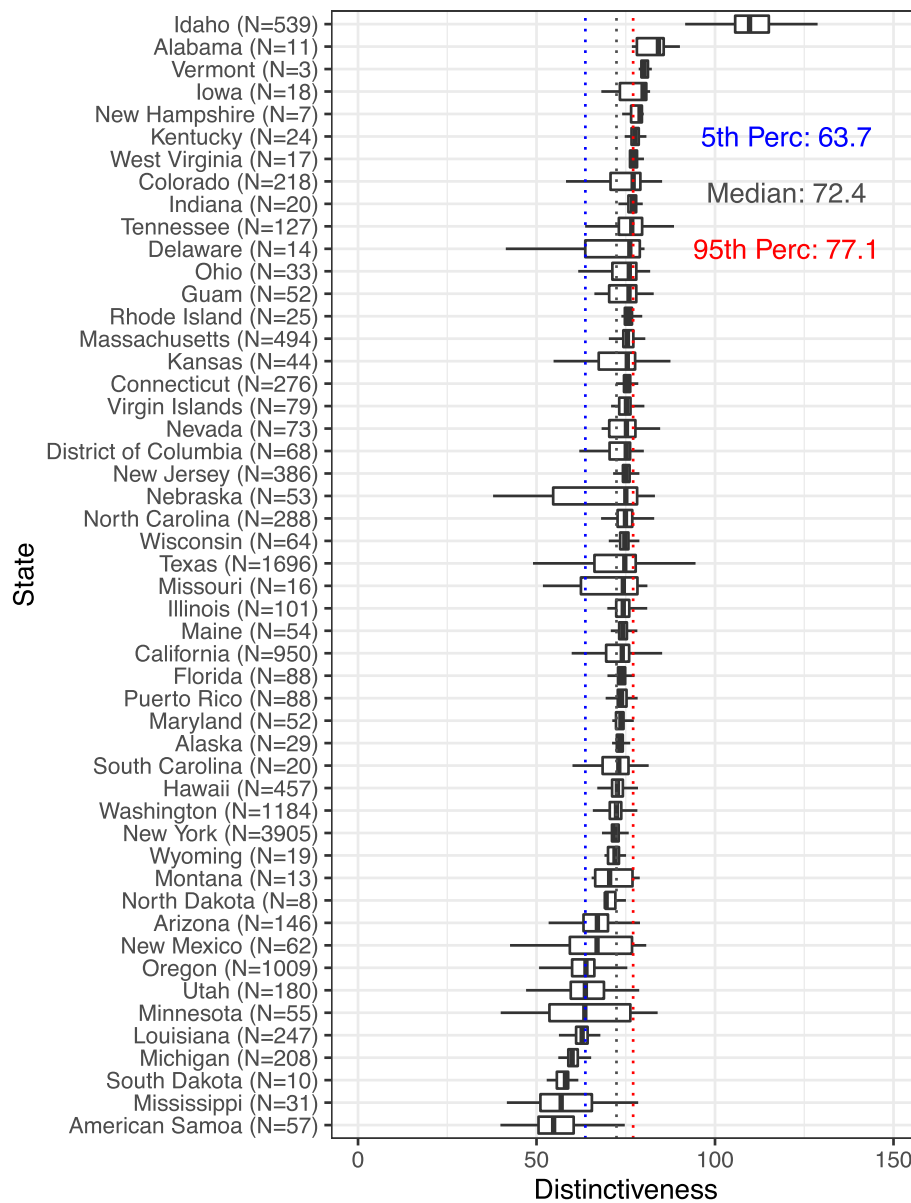


Fig. 4. Distributions of Distinctiveness of Omicron sequences within US states after 2021 November 30.

Discussion

The Distinctiveness metric defined here provides an intuitive quantification of the extent to which any viral sequence differs from other sequences that circulated previously, within the same geographical region. As such, it captures both the emergence of new amino-acid substitutions (e.g. D614G) (23) and deletion of sequence regions that may be involved in antibody recognition (16, 24), both of which can affect viral sequence immunogenicity and infectivity. We find that Distinctiveness can predict future changes in local prevalence of newly circulating lineages, suggesting that Distinctiveness could contribute to accurate and early identification of newly circulating lineages that are likely to outcompete other contemporary lineages. For example, analyzing diversity in the Distinctiveness at the US state-level for the Omicron variant, there are high Distinctiveness sequences in Idaho (Figure 4), warranting a future investigation of sub-regional Distinctiveness within variants and their determinants.

Host immunity against SARS-CoV-2 is largely derived from two sources: vaccination and prior infection. All authorized COVID-19 vaccines utilize the Spike protein sequence from the ancestral Wuhan strain, with a slight modification (substitution of two prolines at positions 986 to 987) to stabilize the prefusion state of the protein product. These vaccines have demonstrated high effectiveness in clinical trials and various real-world studies (17, 25–40), including against most VOCs with the notable recent exception of reduced effectiveness against the Omicron variant (41, 42). With over 10 billion vaccine doses administered around the world, it is likely that vaccination-elicited immunity (i.e. antibody and T cell responses against the ancestral Spike protein sequence) acts as a considerable evolutionary pressure on SARS-CoV-2 (1). The importance of natural immunity as an evolutionary pressure is highlighted by several recent studies demonstrating that prior infection confers robust and durable protection against future infection (3–12). Furthermore, the approach described here can be readily extended to include a correction for the durability

of immunity, for example, by reduced contributions to the Distinctiveness calculation of sequences based on their collection date. We suggest that any newly emerging lineage with a combination of sequence modifications that distinguish it from the ancestral strain and VOCs that have circulated widely (or at high prevalence in a given geographic region) should be monitored closely for their potential to drive future surges.

This study has a few limitations. First, we emphasize that the Distinctiveness metric is intended as an intuitive initial evaluation of the novelty of sampled SARS-CoV-2 sequences. By design, it provides a quantification of prior herd exposure, which is a key contributor to population level immunity. However, there are additional factors, such as the functional implications of mutations and stochasticity, that determine whether a new lineage will spread widely that are not captured by our metric and that can only be captured by subsequent lab assays or epidemiological studies. Future work could combine Distinctiveness with such additional contributing factors to develop a more robust predictor of lineage epidemiological fitness. Second, SARS-CoV-2 genomic epidemiology is unfortunately impacted by major geographic and temporal sequencing biases. Over 55% of SARS-CoV-2 genome sequences in GISAID were isolated from infected patients in the United States or the United Kingdom, and the number of cases subjected to whole genome sequencing increased massively starting at the end of 2020. Undersampling of SARS-CoV-2 genomes in other regions and/or during earlier months of the pandemic could impact our estimations of lineage Distinctiveness. Future analysis will include SARS-CoV-2 genomes from complementary databases such as the National Center for Biotechnology Information (43). Third, although we suggest a cut-off for Distinctiveness that can be used to monitor future emerging lineages (Figure 3), it is not clear that this cut-off will remain optimal. The future of SARS-CoV-2 evolution is uncertain, and may involve smaller changes to the sequence that necessitate a lower cut-off in Distinctiveness, or a more sensitive method, such as one focussed only on key immunogenic positions. Fourth, Distinctiveness can be sensitive to sequence alignment parameters. Complementary analyses that are independent of sequence alignments are warranted to overcome this shortcoming (44). Finally, Distinctiveness does not take into account amino acid similarities in the sequence alignments or the recency of the SARS-CoV-2 sequences used to build the alignment. Future work should account for amino acid similarities using substitution matrices (45) and incorporate the time of sequencing as parameters in computing the Distinctiveness scores.

In conclusion, we highlight that Distinctiveness more holistically captures the ongoing combat between viral evolution and host immunity, wherein lineages which are most distinctive from both the ancestral strain (the basis for all authorized COVID-19 vaccines) and VOCs (i.e. prior dominant strains against which natural immunity has developed) are the least likely to be neutralized by host immune responses. Distinctiveness can be considered as one important feature contributing to the epidemiological fitness of emerging SARS-CoV-2 variants, and thus, a salient factor to monitor as part of the global pandemic preparedness efforts.

Methods

Quantification of number of distinct positional amino acids for prevalent SARS-CoV-2 lineages

Individual substitutions, insertions, and deletions for each aligned SARS-CoV-2 protein sequence along with the corresponding

PANGO designation were obtained from the GISAID (<https://www.gisaid.org>) database, on 2022 May 3. We considered only sequences labeled as “complete” and “high coverage” from the GISAID data, and collected from 28 top sequencing countries (Supplementary Table S1), this resulted in a total of 4,926,906 sequences. For the original, Wuhan strain and the five VOCs (Alpha, Beta, Gamma, Delta, and Omicron), the PANGO classification was obtained from the CDC website (<https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html>).

Calculation of sequence Distinctiveness

For a given sequence, Distinctiveness within a geographical region of interest (i.e. a country) is defined as the average distances at the amino-acid level between that sequence and all sequences that were collected at least one calendar day before that sequence (limited by the time-resolution of the data). Specifically, for a sequence, s , its Distinctiveness, $D(s)$, is calculated using the following formula

$$D(s) = \frac{1}{N_p} \sum_{s' \in \text{prior sequences}} \sum_{p=1}^{\#AA} 1 - \delta(s(p) - s'(p)),$$

where N_p is the number of prior sequences, s' is one specific prior sequence, the inner sum is over all pairwise aligned amino acid positions, and $\delta(s(p) - s'(p))$ evaluates to 1 if sequence s and s' have the same amino-acid identity (one of 20 amino acids, a deletion, or a specific insertion) at position p and 0 otherwise. Positions of amino acids are determined relative to the Wuhan-Hu-1 reference, and insertions were treated as a single modification at the site of insertion. In cases where a nonsense mutation occurred, resulting in an early stop codon, mutations that followed this stop codon were not considered.

Calculation of sequence mutational load

The mutational load was calculated as the number of mutations away from the ancestral Wuhan-Hu-1 sequence. Similar to in the Distinctiveness calculation, insertions were counted as a single mutation. In cases where a nonsense mutation occurred, resulting in an early stop codon, mutations that followed this stop codon were not considered.

Calculating local prevalence of VOCs

The local prevalence of a SARS-CoV-2 variant, as reported in Figure 2 was calculated as the percentage of SARS-CoV-2 sequences in GISAID that were assigned to a lineage comprising that variant, during specific time windows and in specific countries.

Correlating the Distinctiveness and changes in future prevalence of SARS-CoV-2 lineages

We correlated the average Distinctiveness of sequences in a set during a 28-d window to the change in prevalence of the corresponding set, defined as prevalence ($t + 56$ to $t + 84$)—prevalence (t to $t + 28$), where t denotes time. For the analysis in Figure 3a, we show data points only for time periods in which one of the VOCs (Alpha, Beta, Gamma, Delta, and Omicron) first reached >5% prevalence in a given geographic region (defined as a country or US state); all variants present in the geographical region at included time windows are shown. This results in 944 data points, spanning 154 time windows in 78 geographical regions. An alternate version of this analysis, with inclusion of all available time windows (36,000 time windows spanning the same 78 geographical

regions) is shown in Supplementary Figure S4 and yields similar conclusions as those described in the main text.

ROC-curves were generated from these data using Scikit-learn, using binary labels based on a minimum 20 percentage point increase in lineage prevalence for a country/time datapoint. Resulting AUC and threshold values, maximizing the sum of Sensitivity and Specificity, were found to be robust with respect to the cut-off used for labeling the data based on the percentage point increase (Supplementary Figure S5). We used bootstrap resampling (10,000 samples) of the underlying data points (scatter points in Figure 3a) to estimate 95% CIs on the resulting AUC values.

Labeling of neutralizing antibody epitope sites on the Spike protein

We have abstracted antibody epitope data for Therapeutic antibodies, as tracked by NCATS (<https://opendata.ncats.nih.gov/covid19/>), as well as Neutralizing antibodies, typically isolated from convalescent patient sera, as encountered in the Protein Data Bank (46). We define an antibody epitope as all residues in the antigen protein that have heavy (non-hydrogen) atoms at a distance of 4 Å or less to heavy atoms of the bound antibody. When a structure of an antigen-antibody complex contains multiple instances of the interaction, such as in the case of a Spike protein trimer, and/or when several structures of the same antigen-antibody complex are available, we aggregate the binding data into a single epitope definition. We have also collected data for neutralizing antibodies as listed in Supplementary Data files provided by the Bloom and Xie labs (47, 48), who have reported the results of single-point mutations that affect binding affinities (https://media.githubusercontent.com/media/jbloomlab/SARS_2_RBD_Ab_escape_maps/main/processed_data/escape_data.csv). We have listed residues whose mutations were found to have a nontrivial effect on binding activity for a given antibody (site total escape of 0.1 or higher). These are not necessarily close in 3D structure. As structures of those antibodies with bound antigen become available, we do find good agreement, in general, and we amend the epitope definition with that derived from the 3D structure data. In a few instances, structure-derived epitopes were slightly extended based on the characterization of the epitope by the structure's authors, and may include interactions slightly beyond the 4 Å cutoff that we have employed.

Specifically, the following positions in the Spike protein were labeled neutralizing antibody epitope sites: 13, 14, 19, 64, 66, 67, 69, 70, 75, 76, 77, 126, 140, 142, 143, 144, 145, 146, 148, 152, 153, 154, 156, 157, 158, 211, 212, 213, 243, 244, 245, 250, 251, 253, 258, 262, 346, 367, 373, 375, 376, 394, 405, 408, 410, 411, 414, 417, 440, 446, 449, 450, 452, 477, 484, 486, 489, 490, 493, 494, 496, 498, 505, 562, 1146, and 1147.

Supplementary material

Supplementary material is available at [PNAS Nexus](#) online.

Funding

This study was self-funded by inference. No external funding was received for this study. The work of A.M.B., J.W., R.C., and J.B. was supported by the National Center for Biotechnology Information of the National Library of Medicine, National Institutes of Health.

Authors' contributions

M.N. and K.M. designed research, performed research, analyzed data, and wrote the paper; P.L. designed research, analyzed data, and wrote the paper; A.M.-B. and J.W. wrote the paper; R.C. contributed new reagents/analytic tools, and wrote the paper; J.R.B. wrote the paper; A.J.V. designed research; analyzed data, and wrote the paper; and V.S. designed research, contributed new reagents/analytic tools, and wrote the paper.

Data availability

All SARS-CoV-2 sequences and associated metadata were downloaded from GISAID (<https://www.gisaid.org/>).

References

1. COVID-19 Map. Johns Hopkins Coronavirus Resource Center. <https://coronavirus.jhu.edu/map.html>, (Last accessed: March 10, 2022).
2. CDC. 2020. COVID Data Tracker. Centers for Disease Control and Prevention. <https://covid.cdc.gov/covid-data-tracker/#data-tracker-home>, (Last accessed: March 8, 2022).
3. Goldberg Y, et al. 2022. American Journal of Epidemiology. Similarity of Protection Conferred by Previous SARS-CoV-2 Infection and by BNT162b2 Vaccine: A 3-Month Nationwide Experience From Israel. <https://doi.org/10.1093/aje/kwac060>.
4. Shenai MB, Rahme R, Noorchashm H. 2021. Equivalency of Protection From Natural Immunity in COVID-19 Recovered Versus Fully Vaccinated Persons: A Systematic Review and Pooled Analysis. *Cureus* 13, (10): e19102. [10.7759/cureus.19102](https://doi.org/10.7759/cureus.19102).
5. Shrestha NK, Burke PC, Nowacki AS, Terpeluk P, Gordon SM. 2022. Necessity of coronavirus disease 2019 (COVID-19) vaccination in persons who have already had COVID-19. *Clin Infect Dis*. Jan 13:ciac022. [doi:10.1093/cid/ciac022](https://doi.org/10.1093/cid/ciac022).
6. Lumley SF, et al. 2021. An observational cohort study on the incidence of SARS-CoV-2 infection and B.1.1.7 variant infection in healthcare workers by antibody and vaccination status. *Clin Infect Dis* 74:1208–1219. [doi:10.1093/cid/ciab608](https://doi.org/10.1093/cid/ciab608).
7. Cavanaugh AM, Spicer KB, Thoroughman D, Glick C, Winter K. 2021. Reduced risk of reinfection with SARS-CoV-2 After COVID-19 vaccination—Kentucky, May–June 2021. *MMWR Morb Mortal Wkly Rep* 70: 1081–1083.
8. Gazit S, et al. 2022. The Incidence of SARS-CoV-2 reinfection in persons with naturally acquired immunity with and without subsequent receipt of a single dose of BNT162b2 vaccine. *Ann Intern Med* 175:674–681.
9. Charmetant X, et al. 2022. Infection or a third dose of mRNA vaccine elicit neutralizing antibody responses against SARS-CoV-2 in kidney transplant recipients. *Sci Transl Med* 14. [doi:10.1126/scitranslmed.abl6141](https://doi.org/10.1126/scitranslmed.abl6141).
10. León TM. 2022. COVID-19 cases and hospitalizations by COVID-19 vaccination status and previous COVID-19 diagnosis—California and New York, May–November 2021. *MMWR Morb Mortal Wkly Rep* 71:125–131.
11. Chemaitelly H, Bertollini R, Abu-Raddad LJ. 2021. Efficacy of natural immunity against SARS-CoV-2 reinfection with the Beta variant. *N Engl J Med* 385:2585–2586.
12. Hall V, et al. 2022. Protection against SARS-CoV-2 after Covid-19 vaccination and previous infection. *N Engl J Med* 386: 1207–1220.
13. CDC. 2022. SARS-CoV-2 variant classifications and definitions. Centers for Disease Control and Prevention.

- <https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html>. (Last accessed: May 11, 2022).
14. Gobeil SM-C, et al. 2021. Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. *Science* 373: eabi6226. doi:10.1126/science.abi6226.
 15. Plante JA, et al. 2020. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 592:116–121.
 16. McCarthy KR, et al. 2021. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* 371:1139–1142.
 17. Pegu A, et al. 2021. Durability of mRNA-1273 vaccine-induced antibodies against SARS-CoV-2 variants. *Science* 373: 1372–1377.
 18. Wang P, et al. 2021. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* 593: 130–135. doi:10.1038/s41586-021-03398-2.
 19. Collier DA, et al. 2021. Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* 593:136–141.
 20. Alter G, et al. 2021. Immunogenicity of Ad26.COV2.S vaccine against SARS-CoV-2 variants in humans. *Nature* 596: 268–272.
 21. Korber B, et al. 2020. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182:812–827.e19
 22. Yurkovetskiy L, et al. 2020. Structural and functional analysis of the D614G SARS-CoV-2 Spike protein variant. *Cell* 183:739–751.e8.
 23. Zhang L, et al. 2020. SARS-CoV-2 Spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun* 11: 6013.
 24. Venkatakrishnan AJ, et al. 2021. Antigenic minimalism of SARS-CoV-2 is linked to surges in COVID-19 community transmission and vaccine breakthrough infections. medRxiv. <https://www.medrxiv.org/content/10.1101/2021.05.23.21257668v3>, (Last accessed: March 8, 2022).
 25. Pawlowski C, et al. 2021. FDA-authorized mRNA COVID-19 vaccines are effective per real-world evidence synthesized across a multi-state health system. *Med (N Y)* 2: 979–992.e8. doi:10.1016/j.medj.2021.06.007.
 26. Dagan N, et al. 2021. BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. *N Engl J Med* 384: 1412–1423.
 27. Thompson MG, et al. 2021. Effectiveness of covid-19 vaccines in ambulatory and inpatient care settings. *N Engl J Med* 385:1355–1371.
 28. Pilishvili T, et al. 2021. Effectiveness of mRNA Covid-19 vaccine among U.S. health care personnel. *N Engl J Med* 385:e90.
 29. Chemaitelly H, et al. 2021. mRNA-1273 COVID-19 vaccine effectiveness against the B.1.1.7 and B.1.351 variants and severe COVID-19 disease in Qatar. *Nat Med* 27:1614–1621.
 30. Thompson MG, et al. 2021. Interim estimates of vaccine effectiveness of BNT162b2 and mRNA-1273 COVID-19 vaccines in preventing SARS-CoV-2 infection among health care personnel, first responders, and other essential and frontline workers—eight U.S. locations, December 2020–March 2021. *MMWR Morb Mortal Wkly Rep* 70:495–500.
 31. Butt AA, Omer SB, Yan P, Shaikh OS, Mayr FB. 2021. SARS-CoV-2 vaccine effectiveness in a high-risk national population in a real-world setting. *Ann Intern Med* 174:1404–1408.
 32. Jackson LA, et al. 2020. An mRNA vaccine against SARS-CoV-2—preliminary report. *N Engl J Med* 383:1920–1931.
 33. Tartof SY, et al. 2021. Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA: a retrospective cohort study. *Lancet* 398:1407–1416.
 34. Lopez Bernal J, et al. 2021. Effectiveness of Covid-19 vaccines against the B.1.617.2 (Delta) variant. *N Engl J Med* 385:585–594.
 35. Abu-Raddad LJ, Chemaitelly H, Butt AA, National Study Group for COVID-19 Vaccination. 2021. Effectiveness of the BNT162b2 Covid-19 vaccine against the B.1.1.7 and B.1.351 variants. *N Engl J Med* 385:187–189.
 36. Polack FP, et al. 2020. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 383:2603–2615.
 37. Sadoff J, et al. 2021. Safety and efficacy of single-dose Ad26.COV2.S vaccine against Covid-19. *N Engl J Med* 384:2187–2201.
 38. Corchado-Garcia J, et al. 2021. Analysis of the effectiveness of the Ad26.COV2.S adenoviral vector vaccine for preventing COVID-19. *JAMA Netw Open* 4:e2132540.
 39. Falsey AR, et al. 2021. Phase 3 safety and efficacy of AZD1222 (ChAdOx1 nCoV-19) Covid-19 vaccine. *N Engl J Med* 385:2348–2360.
 40. Voysey M, et al. 2021. Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet* 397:99.
 41. Tseng HF, et al. 2022. Effectiveness of mRNA-1273 against SARS-CoV-2 Omicron and Delta variants. *Nat Med* 28: 1063–1071. doi:10.1038/s41591-022-01753-y.
 42. Dorabawila V, et al. 2022. Effectiveness of the BNT162b2 vaccine among children 5–11 and 12–17 years in New York after the emergence of the Omicron variant. *bioRxiv*. doi:10.1101/2022.02.25.22271454.
 43. NCBI SARS-CoV-2 Resources. <https://www.ncbi.nlm.nih.gov/sars-cov-2/>, (Last accessed: June 21, 2022).
 44. Murugadoss K, et al. 2022. Continuous genomic diversification of long polynucleotide fragments drives the emergence of new SARS-CoV-2 variants of concern. *PNAS Nexus* 1: pgac018.
 45. Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89:10915–10919.
 46. Berman H, Henrick K, Nakamura H. 2003. Announcing the worldwide Protein Data Bank. *Nat Struct Mol Biol* 10:980–980.
 47. Cao Y, et al. 2022. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* 602:657–663.
 48. Greaney AJ, et al. 2021. Complete mapping of mutations to the SARS-CoV-2 Spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* 29:44–57 e9.