Competitive fitness of emerging SARS-CoV-2 variants is linked to their Distinctiveness relative to preceding lineages from that region

Michiel J.M. Niesen[1+], Karthik Murugadoss[1+], AJ Venkatakrishnan[1], Patrick J. Lenehan[1], Venky Soundararajan[1*]

[1] nference, Cambridge, Massachusetts 02139, USA

+ Joint first authors

*Correspondence to: Venky Soundararajan (venky@nference.net)

Abstract

The COVID-19 pandemic has seen the persistent emergence of fitter Variants of Concern (VOCs) that have successfully out-competed circulating strains, but the determinants of viral fitness remain unknown. Here we define 'Distinctiveness' of SARS-CoV-2 sequences based on a proteome-wide comparison with all prior sequences from the same geographical region. From the perspective of viral evolution, Distinctiveness captures "regional herd exposure" and has the advantage over the canonical concept of mutation, which relies foremost on the reference ancestral sequence that is invariant over time. By assessing the correlation between Distinctiveness and change in prevalence for all circulating lineages in each region when a new lineage is introduced, we find that the relative Distinctiveness of emergent SARS-CoV-2 lineages is associated with their competitive fitness (Pearson r = 0.67). Further, 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, analysis of Omicron lineages in India and USA shows the BA.1 and BA.2 sub-lineages have comparable distinctiveness, suggesting that they may have similar levels of competitive fitness. Overall, our study proposes that augmenting the ongoing surveillance of highly mutated variants with real-time assessment of Distinctiveness can aid in achieving robust pandemic preparedness.

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 also 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.[13] 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 lineages evolve, understanding the determinants of fitter strains and detecting potential Variants of Concern early is imperative.

With this context, we reasoned that comparing new SARS-CoV-2 lineages with the ancestral strain and previously circulating strains may reflect its likelihood of evading existing immunity and transmitting highly at the community level. We recently found that the genomes of successive VOCs tended to be more distinctive from each other as assessed through the lens of various length polynucleotides.[14] This polynucleotide Distinctiveness metric distinguished VOCs more robustly than various standard phylogenetic distance metrics. Since the primary known sources of immunologic selection pressure (antibodies and T cells) recognize protein sequences, we aimed to determine whether a similar pattern holds for the SARS-CoV-2 peptidome.

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. Rather than simply considering the conventionally defined mutations relative to the ancestral strain, 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 find that the relative Distinctiveness of emergent SARS-CoV-2 lineages is associated with their competitive fitness, as defined by the change in the lineage prevalence. Finally, we show that the same lineage can have different Distinctiveness-contributing positions in different countries.

Results

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

Given the urgent need for early identification of fitter SARS-CoV-2 variants, a robust metric must: (i) capture the novelty of a new sequence by accounting for the entirety of SARS-CoV-2 evolution till-date instead of relying on the ancestral sequence as an unchanging reference and (ii) 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 **Methods**). 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.

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

We computed mutational load and Distinctiveness during the emergence of the VOCs in the country of their emergence. Both mutational load and Distinctiveness were significantly higher

than contemporary lineages (**Figures S1,2**). 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 was significantly higher. Indeed, the Delta variant outcompeted the Gamma variant to become the dominant strain in Brazil (**Figure 2b**). In order to examine whether this trend was generalizable globally, we assessed the correlation between Distinctiveness and change in prevalence for all circulating lineages in 28 countries. We find that the relative Distinctiveness of emergent SARS-CoV-2 lineages is associated with their competitive fitness (Pearson r = 0.67), defined as the change in lineage prevalence over eight weeks (**Figure 2c, Figure S3**). In comparison, mutational load has a lower association with competitive fitness (Pearson r = 0.41).

Given the recent spread of Omicron, we analyzed the Distinctiveness of Omicron (BA.1 and BA.2 lineages) using the Omicron sequences from India and USA as examples. As expected, the Distinctiveness of Omicron lineages is significantly higher than contemporary sequences (**Figure 3**). Further, it is interesting to note that the Distinctiveness values of the Omicron BA.1 and BA.2 sub-lineages are similar (**Figure 3a**), suggesting that they may have similar levels of competitiveness. Also, within the country, there is diversity in the Distinctiveness state-level for the Omicron variant, as observed in the US with high Distinctiveness sequences in Idaho (**Figure 3b**), warranting future investigation of sub-regional Distinctiveness within variants and their determinants.

Same variant can have different Distinctiveness-contributing positions in different countries

Compared to the conventional definition of mutations, Distinctiveness has the intentional advantage of considering previous local herd exposure when evaluating a new lineage. As a result, while mutated positions are fixed based on sequence alignment to the ancestral strain, the positions that contribute to the Distinctiveness of any given viral proteome in two or more geographical regions can vary depending on the prior sequences collected in those regions. To demonstrate this, 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 4a,b**). 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.[15–17] 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 (**Figure 4a,c**), there were 11 other positions that further contributed to its regional Distinctiveness (**Figure 4b,c**). Interestingly, these additional positions correspond to known Gamma lineage-defining mutations (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I, V1176F). This illustrates how Distinctiveness intrinsically accounts for prior herd exposure, as it

effectively compared the Delta variant proteome in Brazil to that of the previously dominant Gamma variant rather than simply defining its features relative to the ancestral strain.

Discussion

Distinctiveness can be considered from at least two complementary angles. First, higher Distinctiveness reflects the acquisition of new amino acid content compared to prior strains, which may confer some evolutionary benefit at the level of infection or replication. For example, when the Spike D614G mutation was first acquired, this would have represented new sequence content (compared to the ancestral strain) that increases infectivity.[18] Further, by definition, any in-frame genomic insertions also generate distinctive amino acid content. On the other hand, high Distinctiveness also implies the modification or loss of amino acid content that was present in one or more previously circulating strains. Teleologically, this would reflect viral evolution to avoid or discard unnecessary or deleterious sequence content, such as sequences that are recognized by host antibodies. Perhaps the most obvious and striking examples of such “sequence loss” for SARS-CoV-2 are the in-frame deletions in the Spike protein N-terminal domain (NTD) which cluster around known binding sites for neutralizing antibodies.[19,20]

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 986-987) to stabilize the pre-fusion state of the protein product. These vaccines have demonstrated high effectiveness in clinical trials and various real-world studies,[21–37] including against most VOCs with the notable recent exception of reduced effectiveness against the Omicron variant.[38,39] 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.[40] 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] 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, 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[41]. Second, it is not yet clear whether there exists a specific threshold for Distinctiveness (or change in Distinctiveness) that should be considered in the monitoring of future emerging lineages. Our retrospective observations show that sequential VOCs harbor progressively more distinctive amino acid content and are more distinctive than other lineages that were in circulation around their time of

emergence, but it is worthwhile to continue prospectively investigating whether a particular degree of increased Distinctiveness is necessary for a new lineage to effectively spread within a region or across the globe. Third, Distinctiveness can be sensitive to sequence alignment parameters. Complementary analyses that are independent of sequence alignments are warranted to overcome this shortcoming. 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[42] 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 competitive 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 sequence along with the corresponding PANGO designation were obtained from the GISAID (https://www.gisaid.org) database. Unless otherwise indicated, we considered only sequences labeled as “complete” and “high coverage” from the GISAID data. Only in the analysis presented in **Figure 3**, focused on the Omicron lineages, the “high coverage” filter was dropped, as this filter led to the exclusion of ~97% of complete Omicron sequences (compared with 27% for all other lineages). 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*, it’s 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 twenty 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 variants of concern

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 competitiveness of SARS-CoV-2 lineages

We correlated the average Distinctiveness of sequences in a set during a 28 day 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. Only countries with at least 100 sequences collected in each of the two 28-day time windows were considered. For the analysis in **Figure 2C** 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 country; all variants present in the country at included time windows are shown. This results in 280 data points, spanning 71 time windows in 28 countries. An alternate version of this analysis, with inclusion of all available time windows (1,511 time windows) is shown in **Figure S3** and yields similar conclusions as those described in the main text.

Declaration of Interests

All authors 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.

Data Availability

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

Funding Statement

This study was self-funded by nference. No external funding was received for this study.

References

1. COVID-19 Map. *Johns Hopkins Coronavirus Resource Center* https://coronavirus.jhu.edu/map.html.
2. CDC. COVID Data Tracker. *Centers for Disease Control and Prevention* https://covid.cdc.gov/covid-data-tracker (2020).
3. Goldberg, Y. *et al.* Protection of previous SARS-CoV-2 infection is similar to that of BNT162b2 vaccine protection: A three-month nationwide experience from Israel. *medRxiv* 2021.04.20.21255670 (2021).
4. Shenai, M. B., Rahme, R. & Noorchashm, H. Equivalency of Protection from Natural Immunity in COVID-19 Recovered Versus Fully Vaccinated Persons: A Systematic Review and Pooled Analysis. *medRxiv* 2021.09.12.21263461 (2021).
5. Shrestha, N. K., Burke, P. C., Nowacki, A. S., Terpeluk, P. & Gordon, S. M. Necessity of Coronavirus Disease 2019 (COVID-19) Vaccination in Persons Who Have Already Had COVID-19. *Clin. Infect. Dis.* (2022) doi:10.1093/cid/ciac022.
6. Lumley, S. F. *et al.* 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.* (2021) doi:10.1093/cid/ciab608.
7. Cavanaugh, A. M. Reduced Risk of Reinfection with SARS-CoV-2 After COVID-19 Vaccination — Kentucky, May–June 2021. *MMWR Morb. Mortal. Wkly. Rep.* **70**, (2021).
8. Gazit, S. *et al.* 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.* (2022) doi:10.7326/M21-4130.
9. Charmetant, X. *et al.* Infection or a third dose of mRNA vaccine elicit neutralizing antibody responses against SARS-CoV-2 in kidney transplant recipients. *Sci. Transl. Med.* (2022) doi:10.1126/scitranslmed.abl6141.
10. León, T. M. 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**, (2022).
11. Chemaitelly, H., Bertollini, R. & Abu-Raddad, L. J. Efficacy of Natural Immunity against SARS-CoV-2 Reinfection with the Beta Variant. *N. Engl. J. Med.* **385**, (2021).
12. Hall, V. *et al.* Protection against SARS-CoV-2 after Covid-19 Vaccination and Previous Infection. *N. Engl. J. Med.* (2022) doi:10.1056/NEJMoa2118691.
13. CDC. SARS-CoV-2 Variant Classifications and Definitions. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html (2021).
14. Murugadoss, K. *et al.* Genomic Diversification of Long Polynucleotide Fragments Is a Signature of Emerging SARS-CoV-2 Variants of Concern. (2021) doi:10.2139/ssrn.3993373.
15. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116–121 (2020).
16. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020).
17. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183**, 739–751.e8 (2020).
18. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* **11**, (2020).
19. McCarthy, K. R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* **371**, (2021).
20. Venkatakrishnan, A. J. *et al.* Antigenic minimalism of SARS-CoV-2 is linked to surges in COVID-19 community transmission and vaccine breakthrough infections. *medRxiv* 2021.05.23.21257668 (2021).
21. Pawlowski, C. *et al.* FDA-authorized mRNA COVID-19 vaccines are effective per real-world evidence

synthesized across a multi-state health system. *Med (N Y)* (2021) doi:10.1016/j.medj.2021.06.007.
22. Dagan, N. *et al.* BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. *N. Engl. J. Med.* **384**, 1412–1423 (2021).
23. Thompson, M. G. *et al.* Effectiveness of Covid-19 Vaccines in Ambulatory and Inpatient Care Settings. *N. Engl. J. Med.* (2021) doi:10.1056/NEJMoa2110362.
24. Pilishvili, T. *et al.* Effectiveness of mRNA Covid-19 Vaccine among U.S. Health Care Personnel. *N. Engl. J. Med.* (2021) doi:10.1056/NEJMoa2106599.
25. Pegu, A. *et al.* Durability of mRNA-1273 vaccine–induced antibodies against SARS-CoV-2 variants. *Science* (2021) doi:10.1126/science.abj4176.
26. Chemaitelly, H. *et al.* 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.* (2021) doi:10.1038/s41591-021-01446-y.
27. Thompson, M. G. *et al.* 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 (2021).
28. Butt, A. A., Omer, S. B., Yan, P., Shaikh, O. S. & Mayr, F. B. SARS-CoV-2 Vaccine Effectiveness in a High-Risk National Population in a Real-World Setting. *Ann. Intern. Med.* (2021) doi:10.7326/M21-1577.
29. Jackson, L. A. *et al.* An mRNA Vaccine against SARS-CoV-2 - Preliminary Report. *N. Engl. J. Med.* **383**, 1920–1931 (2020).
30. Tartof, S. Y. *et al.* 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* (2021) doi:10.1016/S0140-6736(21)02183-8.
31. Lopez Bernal, J. *et al.* Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *N. Engl. J. Med.* **385**, 585–594 (2021).
32. Abu-Raddad, L. J., Chemaitelly, H., Butt, A. A. & National Study Group for COVID-19 Vaccination. 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 (2021).
33. Polack, F. P. *et al.* Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N. Engl. J. Med.* **383**, (2020).
34. Sadoff, J. *et al.* Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19. *N. Engl. J. Med.* **384**, (2021).
35. Corchado-Garcia, J. *et al.* Analysis of the Effectiveness of the Ad26.COV2.S Adenoviral Vector Vaccine for Preventing COVID-19. *JAMA Netw Open* **4**, e2132540–e2132540 (2021).
36. Falsey, A. R. *et al.* Phase 3 Safety and Efficacy of AZD1222 (ChAdOx1 nCoV-19) Covid-19 Vaccine. *N. Engl. J. Med.* **385**, (2021).
37. Voysey, M. *et al.* 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 (2021).
38. Tseng, H. F. *et al.* Effectiveness of mRNA-1273 against SARS-CoV-2 Omicron and Delta variants. *Nat. Med.* (2022) doi:10.1038/s41591-022-01753-y.
39. Dorabawila, V. *et al.* Effectiveness of the BNT162b2 vaccine among children 5-11 and 12-17 years in New York after the Emergence of the Omicron Variant. *bioRxiv* (2022) doi:10.1101/2022.02.25.22271454.
40. COVID-19 Map. *Johns Hopkins Coronavirus Resource Center* https://coronavirus.jhu.edu/map.html.
41. NCBI SARS-CoV-2 Resources. https://www.ncbi.nlm.nih.gov/sars-cov-2/.
42. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).

Figures

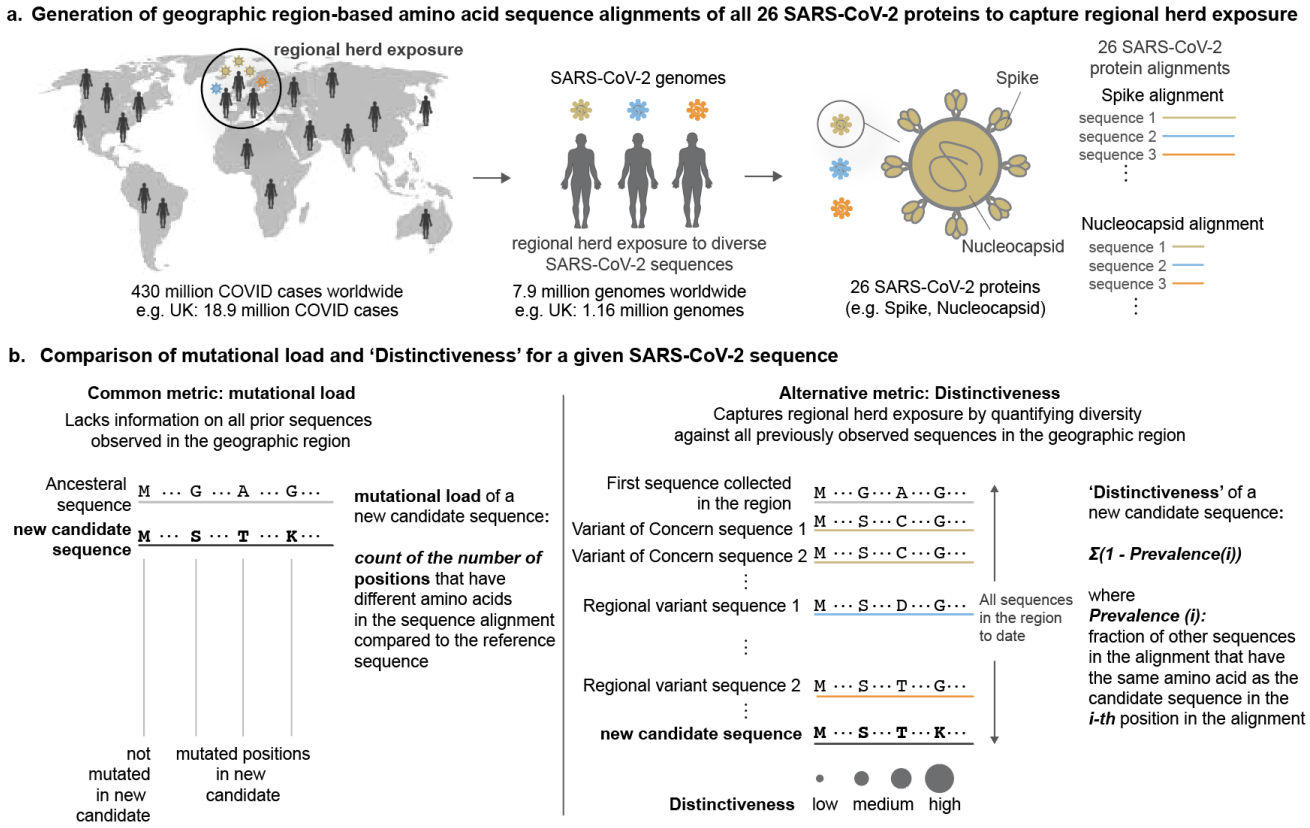

Figure 1. a. Generation of geographic region-based amino acid sequence alignments of all 26 SARS-CoV-2 proteins to capture regional herd exposure. **b.** Comparison of mutational load and 'Distinctiveness' for a given SARS-CoV-2 sequence.

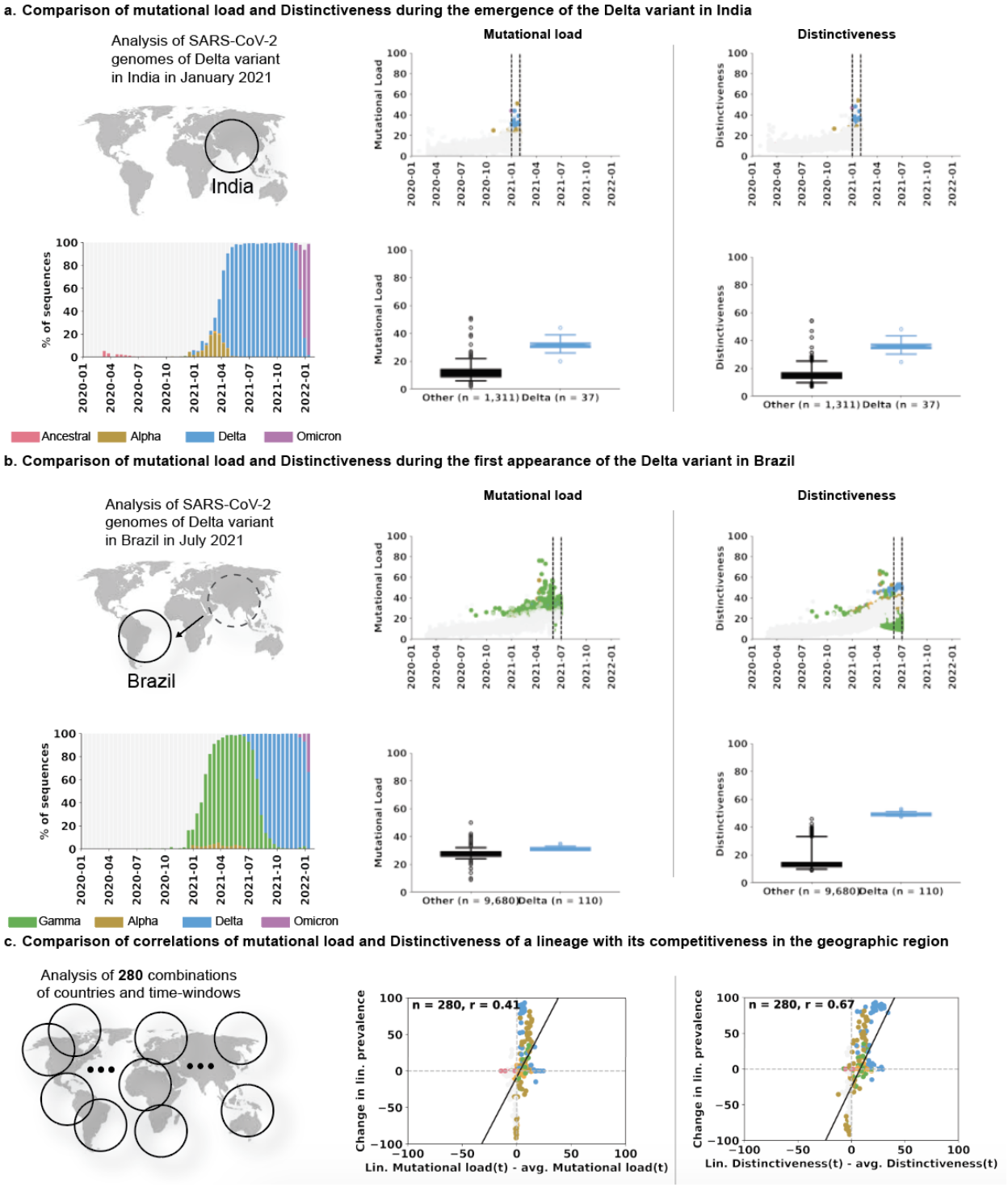

Figure 2: Sequence Distinctiveness as a function of time in geographical regions where VOCs first emerged. Comparison of mutational load and Distinctiveness during the emergence of the Delta variant in: **a.** India, and **b.** Brazil. **c.** Comparing the correlations of mutational load and Distinctiveness of a lineage with its competitiveness in the geographic region (across 71 time windows from 28 countries). **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).

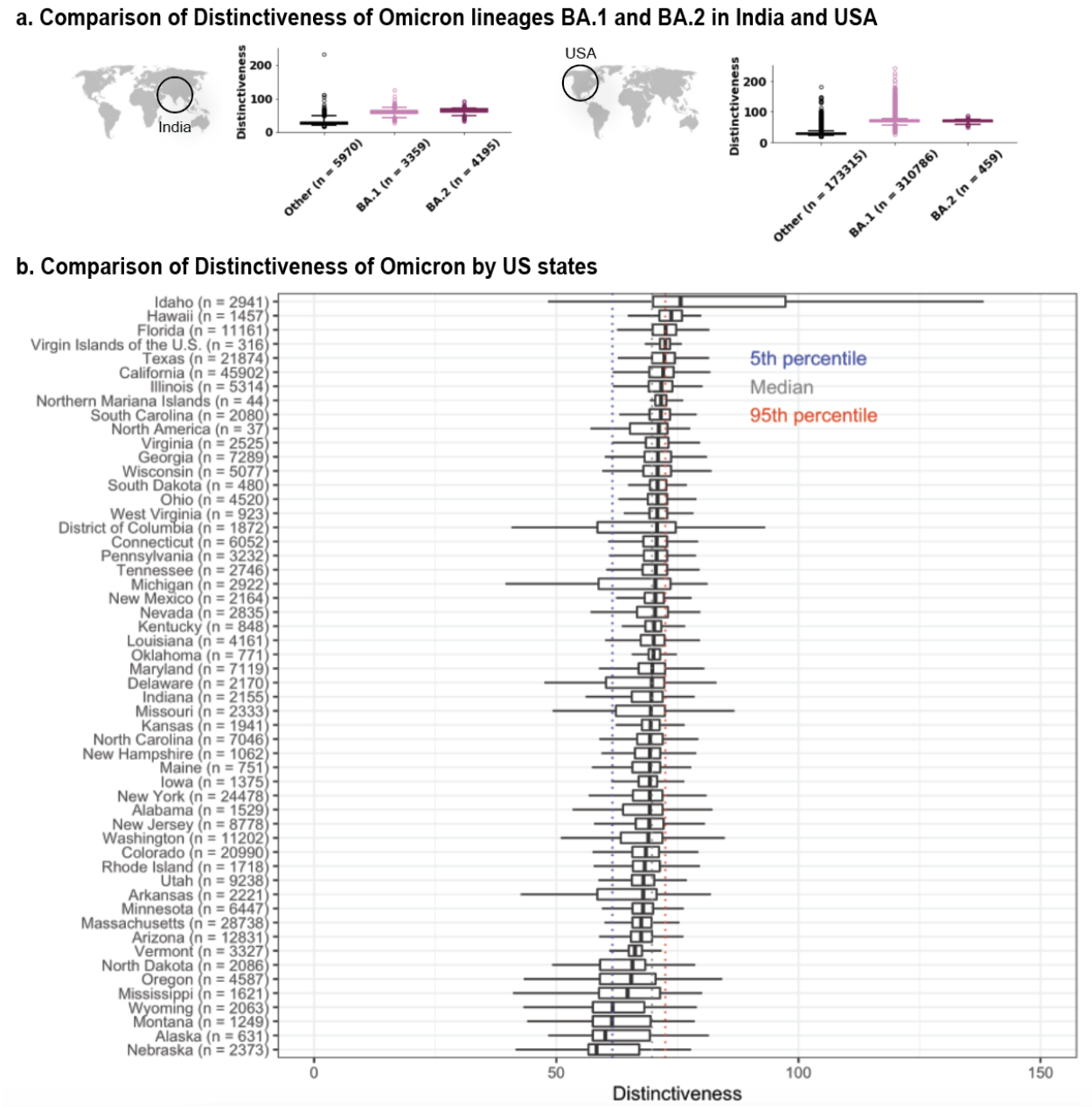

Figure 3: a. Comparison of Distinctiveness between Omicron Lineages BA.1 and BA.2 in India and United States during the period between November 30, 2021 to December 06, 2022 using boxplots **b.** Distributions of Distinctiveness of Omicron sequences within US states during the period between November 30, 2021 to December 06, 2022 using boxplots.

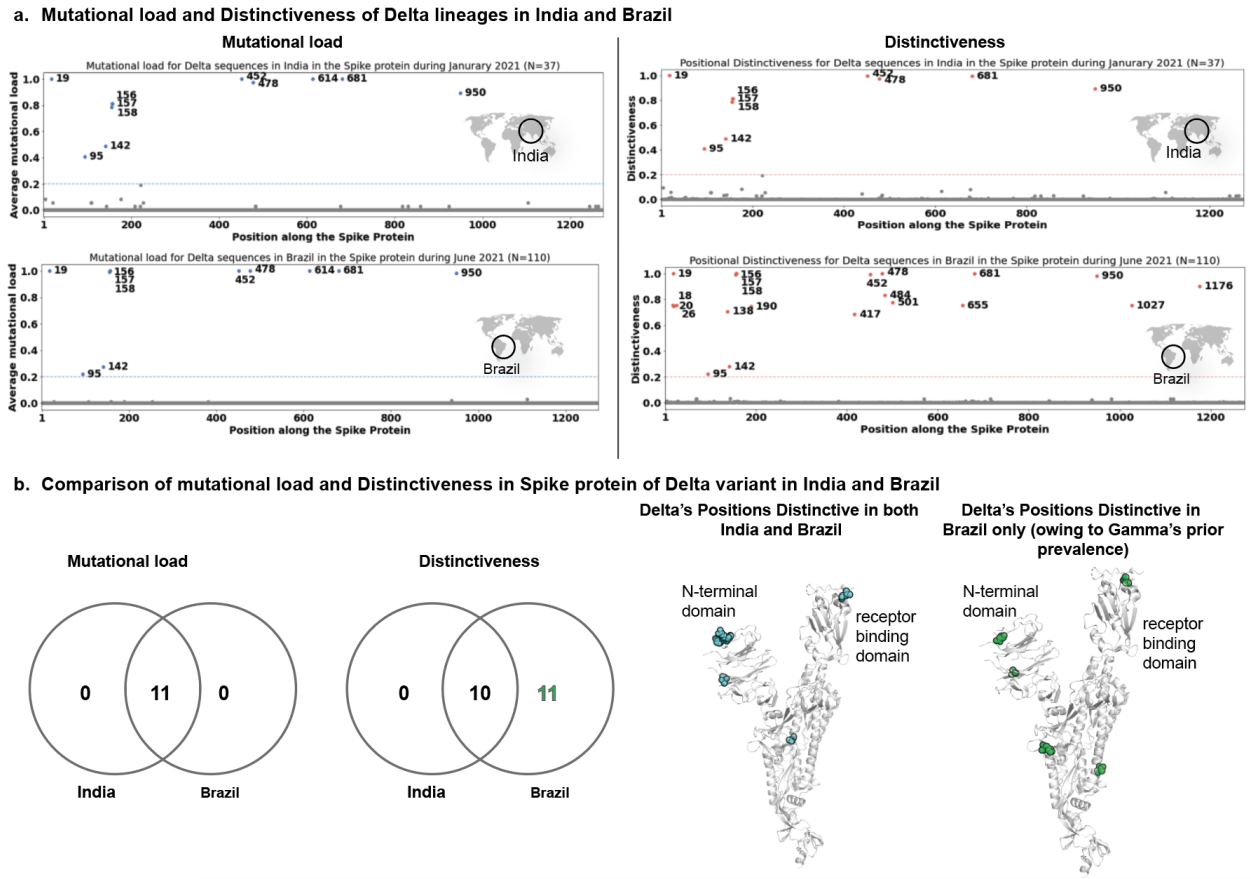

Figure 4: a. Mutational load and Distinctiveness of Delta lineages in India and Brazil. The x-axes denote the amino acid positions in the Spike protein and the y-axes denote the average mutational load (left panel) or the Distinctiveness (right panel). Horizontal lines at y=0.2 denote a high threshold above which amino acid positions are labeled. The world maps are shown as insets and the countries are highlighted using a circle. **b.** 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 Figure 4a). The positions that have high Distinctiveness are highlighted on the protein structures of the Spike protein as spheres (PDB identifier: 6VSB).

Supplementary Information

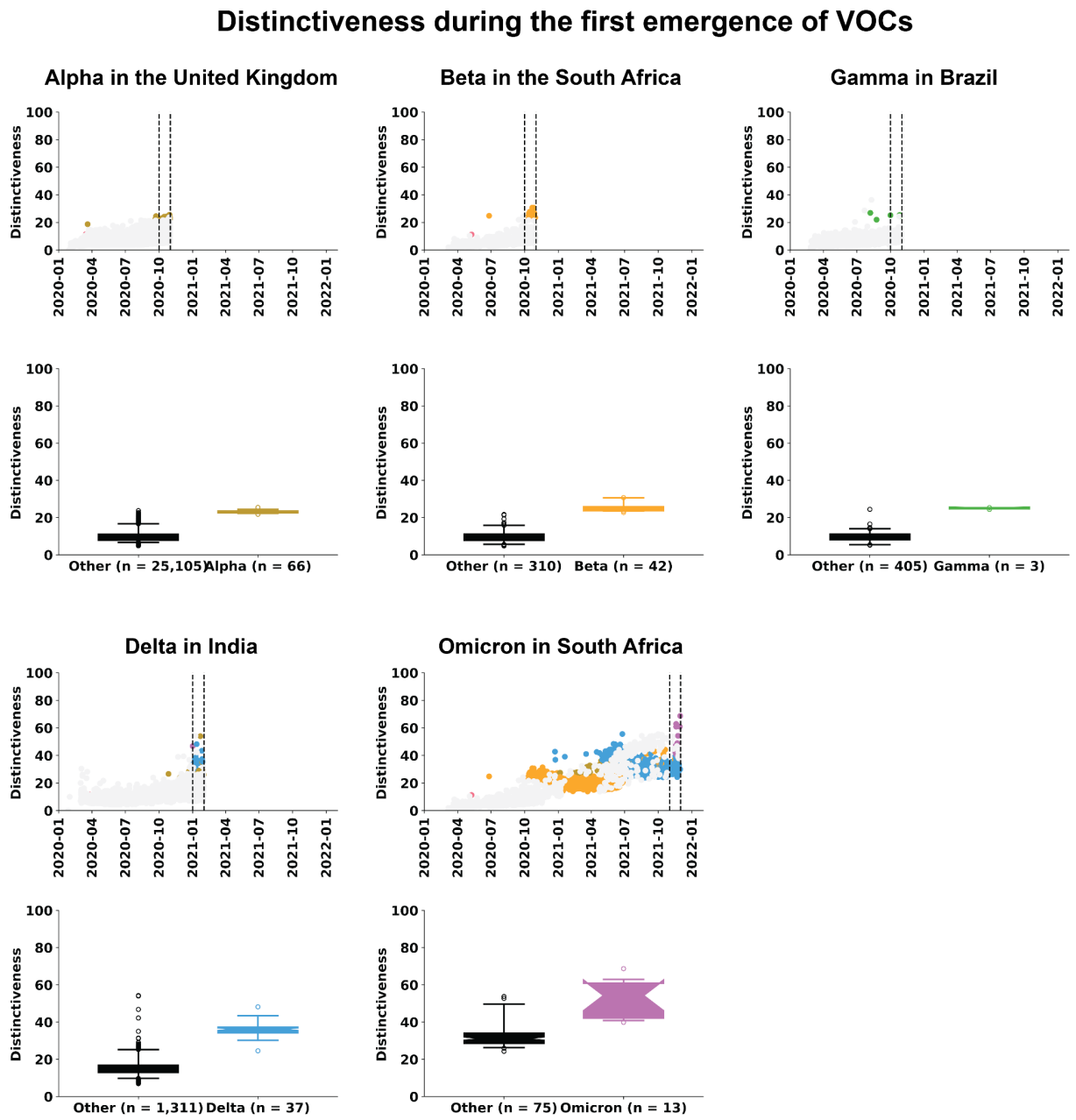

Figure S1: Distinctiveness of variants of concern during the time when they first appeared. In all cases, the Distinctiveness of the VOCs is significantly higher (p-value < 0.001) than that of contemporary sequences. For Alpha, Beta, and Gamma, Distinctiveness values of sequences collected during October 2020 are shown; for Delta Distinctiveness values of sequences collected during January 2021 are shown; and for Omicron Distinctiveness values of sequences collected during November 2021 are shown.

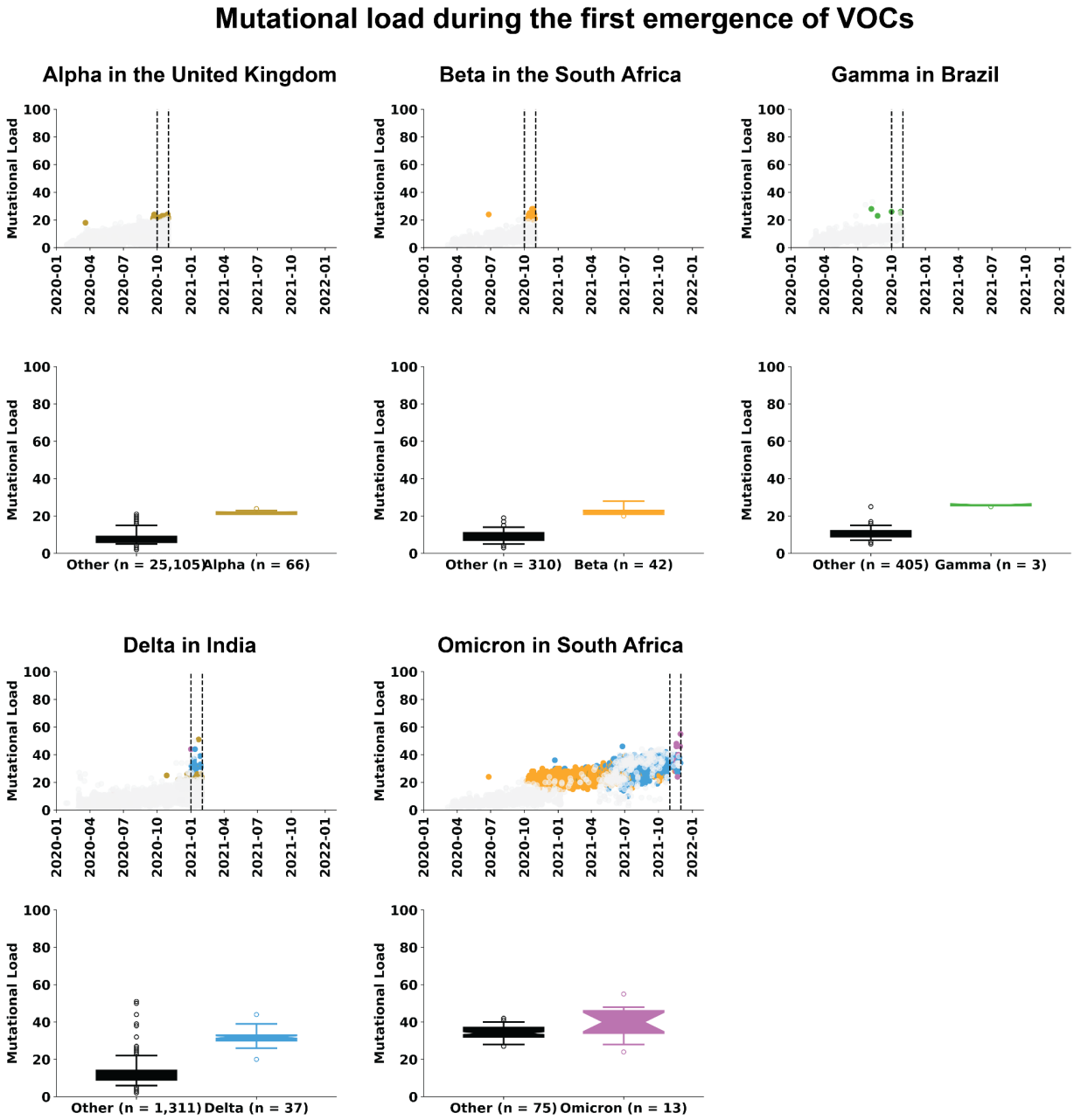

Figure S2: Mutational load of variants of concern during the time when they first appeared. In all cases, the Mutational load of the VOCs is significantly higher (p-value < 0.001) than that of contemporary sequences. For Alpha, Beta, and Gamma, Mutational load values of sequences collected during October 2020 are shown; for Delta Mutational load values of sequences collected during January 2021 are shown; and for Omicron Mutational load values of sequences collected during November 2021 are shown.

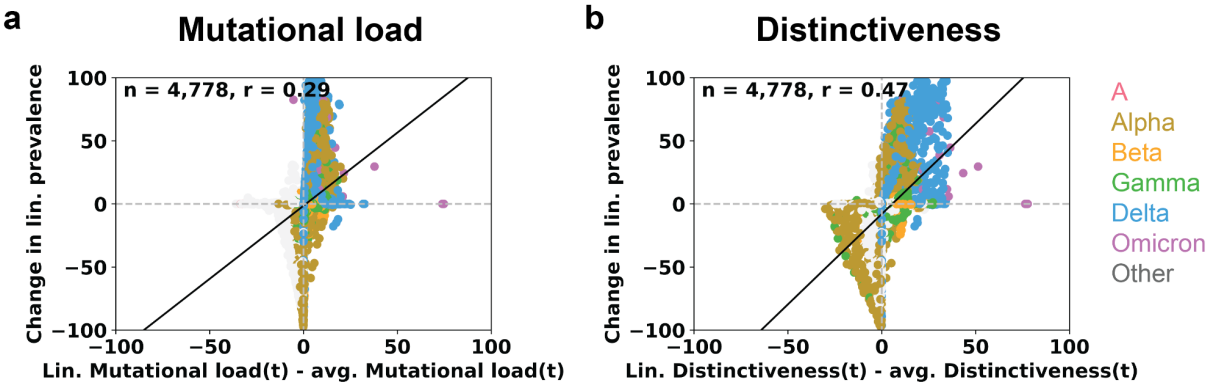

Figure S3: Correlation between the change in prevalence of a lineage, from the current time to 56 days in the future, and: (**a**) the average Mutational load of sequences assigned to that lineage, compared with contemporary sequences, or (**b**) the average Distinctiveness of sequences assigned to that lineage, compared with contemporary sequences. Changes in prevalence and corresponding Mutational load or Distinctiveness values were calculated for 1,511 time intervals from 28 countries.