Continuous genomic diversification of long polynucleotide fragments drives the emergence of new SARS-CoV-2 variants of concern

Karthik Murugadoss\textsuperscript{a,†}, Michel J.M. Niesen\textsuperscript{a,†}, Bharathwaj Raghunathan\textsuperscript{b}, Patrick J. Lenehan\textsuperscript{a,†}, Pritha Ghosh\textsuperscript{c}, Tyler Feener\textsuperscript{a,†}, Praveen Anand\textsuperscript{c}, Safak Simsek\textsuperscript{c}, Rohit Suratekar\textsuperscript{c}, Travis K. Hughes\textsuperscript{a} and Venky Soundararajan\textsuperscript{a,†,∗}

\textsuperscript{a}nference, Cambridge, MA 02139, USA
\textsuperscript{b}nference, Toronto, ON M5V 1M1, Canada
\textsuperscript{c}nference Labs, Bengaluru, Karnataka 560017, India

†Joint first authors
∗To whom correspondence should be addressed: Venky Soundararajan. Email: venky@nference.net

Abstract

Highly transmissible or immuno-evasive SARS-CoV-2 variants have intermittently emerged, resulting in repeated COVID-19 surges. With over 6 million SARS-CoV-2 genomes sequenced, there is unprecedented data to decipher the evolution of fitter SARS-CoV-2 variants. Much attention has been directed to studying the functional importance of specific mutations in the Spike protein, but there is limited knowledge of genomic signatures shared by dominant variants. Here, we introduce a method to quantify the genome-wide distinctiveness of polynucleotide fragments (3- to 240-mers) that constitute SARS-CoV-2 sequences. Compared to standard phylogenetic metrics and mutational load, the new metric provides improved separation between Variants of Concern (VOCs; Reference = 89, IQR: 65–108; Alpha = 166, IQR: 149–181; Beta 131, IQR: 114–149; Gamma = 164, IQR: 150–178; Delta = 235, IQR: 217–255; and Omicron = 459, IQR: 395–521). Omicron’s high genomic distinctiveness may confer an advantage over prior VOCs and the recently emerged and highly mutated B.1.640.2 (IHU) lineage. Evaluation of 883 lineages highlights that genomic distinctiveness has increased over time ($R^2 = 0.37$) and that VOCs score significantly higher than contemporary non-VOC lineages, with Omicron among the most distinctive lineages observed. This study demonstrates the value of characterizing SARS-CoV-2 variants by genome-wide polynucleotide distinctiveness and emphasizes the need to go beyond a narrow set of mutations at known sites on the Spike protein. The consistently higher distinctiveness of each emerging VOC compared to prior VOCs suggests that monitoring of genomic distinctiveness would facilitate rapid assessment of viral fitness.

Significance Statement:

Highly transmissible or immuno-evasive SARS-CoV-2 variants have intermittently emerged and outcompeted previously circulating strains, resulting in repeated COVID-19 surges, reinfections, and vaccine breakthrough infections. Much attention has been directed to studying how specific point mutations in the SARS-CoV-2 Spike protein impact its binding to the ACE2 receptor or viral neutralization by antibodies, but there is limited knowledge of genomic signatures shared primarily by dominant variants. Here, we show that increasing distinctiveness in long stretches of nucleotides across the SARS-CoV-2 genome is a signature of dominant “Variants of Concern.” The methodology we introduce will help peer into the “crystal ball” of SARS-CoV-2 evolution to begin predicting which dominant variants are likely to emerge next, thus helping make pandemic preparedness more proactive.

Introduction

The emergence of several SARS-CoV-2 Variants of Concern (VOCs: Alpha, Beta, Gamma, Delta, and Omicron) over time has resulted in repeated surges of COVID-19 cases, hospitalizations, and deaths around the globe (1). Phylogenetic classification in the Phylogenetic Assignment of Named Global Outbreak Lineages (PANGO) nomenclature shows that these variants have evolved from common ancestors, but none are direct descendants of one another (2). The PANGO lineages corresponding to these VOCs are as follows: Alpha variant (B.1.1.7 and Q lineages), Beta variant (B.1.351 and descendant lineages), Gamma variant (P.1, which is a descendant of B.1.1.28, and descendant lineages), Delta variant (B.1.617.2 and AV lineages), and Omicron variant (B.1.1.529 and BA lineages) (3). All of these variants evolved from the B.1 lineage, while Alpha, Gamma, and Omicron share the B.1.1 as an additional parent lineage. A recent non-VOC, the IHU variant (B.1.640) that was identified in France also captured interest due to the amount and nature of mutations in this variant (4). However, these
phylogenetic classifications do not intuitively describe the degree of distinctiveness between VOCs, nor do they provide concrete insights into the genomic properties of each variant.

SARS-CoV-2, like other viruses, evolves via the introduction of mutations in its genome (5). In some cases, these mutations yield changes in the amino acid sequence of viral proteins. Such mutations can then be positively or negatively selected depending on their impact on various aspects of viral fitness, including transmissibility (e.g., ability to infect and/or replicate in host cells) and immune evasion (e.g., ability to avoid binding by host-derived neutralizing antibodies). It has become clear from global data sharing efforts, particularly the GISAID (https://www.gisaid.org) (6, 7) and NCBI (https://www.ncbi.nlm.nih.gov/sars-cov-2/) databases that mutations in several regions such as the receptor binding domain (RBD) and N-terminal domain (NTD) of the Spike glycoprotein contribute to improved viral fitness (8–14). Significant effort has been devoted to functionally characterizing the mutations that define VOCs (9, 15–17). However, while much attention has been paid to the consequence of individual mutations at the amino acid level, there has been limited focus on how SARS-CoV-2 evolution explores the possibilities of diversifying its “language” at the level of nucleotide sequences (i.e., toward deciphering the “words” constituting polynucleotide sequences from the viral genome). Such an understanding of the language of viral evolution could facilitate multiple dimensions of pandemic preparedness including the development and improvement of diagnostics and vaccines (18).

We hypothesized that the emergence of more transmissible or immune evasive SARS-CoV-2 variants over time is associated with increased genomic distinctiveness from the original strain and from strains that were previously circulating. To assess this hypothesis, here we introduce a new methodology to quantify the number of distinct nucleotide n-mers (of various sizes) in VOCs to estimate the degree of viral evolution. We find that SARS-CoV-2 variants that have emerged later (Delta, B.1.640, and Omicron) indeed tend to harbor more distinct nucleotide n-mers than those that emerged earlier (Alpha, Beta, and Gamma), with the original parent strain (PANGO lineage A) expectedly having the lowest level of n-mer distinctiveness. Although correlated, this striking trend is not attributable solely to overall mutational load, nor is it explained simply by a reduction in genomic distinctiveness of early VOCs due to the emergence of later VOCs. Taken together, this study demonstrates that, over the 2 years of the COVID-19 pandemic (December 2019–January 2022), SARS-CoV-2 variants with increasing nucleotide distinctiveness have evolved over time.

The concept of polynucleotide diversification introduced in this study represents a new method to map epidemiological features of SARS-CoV-2 (e.g., transmissibility and immune evasiveness) to a genomic compass. To facilitate the continued monitoring of SARS-CoV-2 evolution through this lens, we are launching a freely accessible resource for pandemic preparedness with genomic inference (“Pandemic Preparedness GENI”—https://academia.nferx.com/GENI). We encourage further investigation to determine whether genomic distinctiveness can help to predict the likelihood that newly emerging variants will outcompete viral strains that are contemporaneously circulating in the same geographic regions.

**Results**

**SARS-CoV-2 VOCs with increasing genomic distinction evolved over time**

To quantify the genomic distinctiveness of SARS-CoV-2 VOCs with respect to the original strain (3, 19), we compared the number of distinctive nucleotide 9-mers (DN9s) present in the Wuhan (PANGO lineage A; original strain), Alpha, Beta, Gamma, Delta, and Omicron variants (see Methods, Fig 1A). Specifically, we performed 100,000 repetitions of an iterative sampling experiment in which we selected 1 SARS-CoV-2 genome corresponding to each of the variants listed above (Table S1, Supplementary Material). From each genome, we derived the set of unique 9-mers, and then compared these sets against each other. We determined the number of DN9s, which were present in each given lineage but absent in all others. The number of DN9s appeared to correlate with time of emergence, with highest values observed for Omicron (emerged in November 2021) followed by Delta (March 2021), Alpha and Gamma (September and October 2020, respectively), and finally Beta (September 2020; Fig. 1B). Most DN9s are found in the Spike gene, although there is also variation in the distribution of DN9s across the other SARS-CoV-2 genes (Fig. 1C).

Omicron sequences robustly had more DN9s than all other VOCs (Cohen’s D > 3.0 for each comparison), and Delta sequences had more DN9s than all VOCs other than Omicron (Fig. 2A and B). However, Alpha, Beta, and Gamma showed more overlap, with all Cohen’s D values less than 1.5 relative to each other (Fig. 2B). The same pattern was also observed when comparing the DN9 distributions via Jensen–Shannon (JS) divergence (Fig. 2C). Further, similar results were observed for various other tested lengths of polynucleotide such as 15-mers (Figure S1 and Table S2, Supplementary Material). Considering these findings, it is notable that Alpha, Beta, and Gamma emerged at similar times (September–October 2021) and drove surges in largely distinct regions (Alpha in the United States, United Kingdom, and Asia; Beta in South Africa; and Gamma in South America; Figure S2, Supplementary Material). On the other hand, Delta and Omicron rapidly spread across the globe and became the dominant variant in most regions after emerging in April and November 2021, respectively (20).

**Genomic distinctiveness at the polynucleotide level provides significant additional context to existing distance metrics**

We next asked whether the observed patterns in genomic distinctiveness (as captured by the DN9 metric) were captured by other simple or commonly used metrics for distinguishing SARS-CoV-2 variants. Overall mutational load does highlight Omicron as the most highly mutated VOC and is clearly correlated with our metric (Fig. 3A). However, mutational load alone provides poor resolution between Alpha, Beta, Gamma, and Delta (Fig. 3A–C), with mean values of 27, 26, 24, and 32, respectively. This suggests that the number of distinctive n-mers harbored by a VOC contains information beyond the number of mutations away from the original SARS-CoV-2 genome. This point is further demonstrated by the IHU variant (B.1.640.2), a sublineage of B.1.640; that emerged in October 2021 but has not spread rapidly across the globe (4). Despite having a higher mutational load than Delta, these variants have similar distinctiveness at the polynucleotide level (Fig. 4A). Together, these examples show that increasing mutational load does not imply an increase in polynucleotide distinctiveness and vice versa. We have also compared the separate Omicron lineages, BA.1 and BA.2, showing that the BA 2 lineage has a higher polynucleotide distinctiveness (Fig. 4B).

The gene-level summary of the distinctive 9-mers present across the original SARS-CoV-2 strain, the VOCs, the BA.1, the BA.2, and the B.1.640 lineage are provided (Figs 1C and 4C). This shows that considering only the spike (S) protein is insufficient to distinguish...
distinctiveness of the VOCs in a way consistent with their emergence and dominance. For example, while the Delta variant has a higher genome-wide distinctiveness than Alpha, Beta, and Gamma, the distinctiveness values for their respective Spike protein-encoding nucleotide sequences are similar. Comparing the Omicron lineages, BA.1 and BA.2, reveals that although the BA.1 lineage has more distinctive polynucleotides in the Spike and NSP3 proteins, the BA.2 lineage has more distinctive polynucleotides in NSP4, NSP5, NSP9, NSP1, NS6, NS3, and overall. Further, it shows that there are several distinctive polynucleotides harbored in highly mutated variants that have not yet circulated globally, such as the B.1.640 lineage.

We also computed pairwise phylogenetic distances between all VOCs using 4 standard metrics (Tajima–Nei (21), Jukes–Cantor (22), Tamura (23), and Kimura (24)). All metrics other than Jukes–Cantor highlight Gamma as most phylogenetically separated from Alpha and Beta, while Jukes–Cantor more strongly reveals the distance between Delta and all other lineages (Fig. 3D–G). Notably, none of these phylogenetic distance metrics reveal the strong distinction between Omicron and other lineages which is captured by both mutational load and our genomic distinctiveness metric.

The emergence of new VOCs has had minimal impact on the genomic distinctiveness of prior VOCs

New variants can emerge that share genomic features with a given prior variant (e.g. due to convergent evolution). In these cases, nucleotide n-mers that were previously considered distinct for that variant will no longer be considered as such, resulting in the potential tendency of DN9s for any lineage to decrease over time after its emergence. In theory, this could at least partially explain the previously described correlation between time of VOC emergence and genomic distinctiveness. To determine whether this is the case, we performed a modified experiment in which we assessed the impact of Delta on the genomic distinctiveness of Alpha, Beta, and Gamma with respect to the original strain.

Even when comparing only Alpha, Beta, and Gamma genomes deposited before June 2021, the DN9s were similar between these...
3 VOCs (Fig. 5A). When additionally considering the Delta variant and expanding to include genomes deposited before November 2021, there was only a slight leftward shift in the DN9 distributions for Alpha, Beta, and Gamma (Fig. 5B). However, the distinctiveness of the original strain decreases notably from a mean of 109 (IQR: 86–129) to a mean of 89 (IQR: 65–109). As expected, Delta genomes tended to have substantially more DN9s than the other VOCs. Comparison of this data to the initial analysis (Fig. 2A–C) also confirms that the emergence of Omicron did not result in a pronounced leftward shift of the genomic distinctiveness of Delta or the other VOCs. Together, this analysis suggests that not only are newly evolving SARS-CoV-2 variants more genomically distinctive compared to the original strain than prior variants, but they also explore unique nucleotide sequences to generate this distinctiveness.

Increasing distinctiveness within the Delta lineage over time was recently outpaced by the highly distinctive Omicron variant

It is also possible that the correlation between genomic distinctiveness and time of emergence simply reflects an evolutionary clock, i.e., the amount of time between the first detection of SARS-CoV-2 and the emergence of each variant. If this is true, then we would expect to observe the following: (i) genomic distinctiveness within a lineage should increase over time; (ii) genomic distinctiveness between lineages should be similar at any given point in time. To test this hypothesis, we modified our initial analysis to compute polynucleotide distinctiveness distributions for Delta genomes divided into 3 nonoverlapping intervals based on their collection date (April, July, or December 2021). The distribution of Delta genome DN9s indeed shifted right over time (April: mean 202, IQR 193–217; July: mean 222, IQR 209–238; and December: mean 267, IQR 247–283; Fig. 6). This indicates that distinctiveness can indeed increase within a lineage with the passage of evolutionary time. However, it is interesting to note that Omicron genomes do tend to have more DN9s than even contemporary Delta genomes (i.e., collected in December; Fig. 6). This suggests that Omicron may be more distinctive than would be expected from evolutionary time alone.

Polynucleotide diversification increases across SARS-CoV-2 lineages over time

In order to compare all SARS-CoV-2 lineages reported to date (rather than restricting to VOCs), we defined another related metric of polynucleotide distinctiveness, termed $A^*(1-B)$, which captures the average specificity to a given lineage of all unique n-mers present in that lineage (see Methods). In addition, this modified metric takes into consideration the prevalence of a polynucleotide n-mer, both within and outside of a lineage; a variable factor that was not explicitly considered in our previous analysis.
Fig. 3. Mutational load and phylogenetic distance analysis of SARS-CoV-2 VOCs. (A) Density plots showing the distribution of the number of mutations observed in SARS-CoV-2 genomes assigned to each listed lineage in the GISAID database. (B) and (C) Heatmaps showing Cohen’s D and JS divergence values from pairwise comparisons of the distribution of distinctive 9-mer sequences between SARS-CoV-2 VOCs. (D)–(G) Heatmaps showing the mean phylogenetic distance calculated using multiple methods (Tajima–Nei, Tamura, Jukes-Cantor, and Kimura) between SARS-CoV-2 VOCs. Abbreviations: VOC—variant of concern.

Discussion
Summary of results
The analyses described here demonstrate that SARS-CoV-2 has continually evolved into variants with increasing levels of genomic distinctiveness with respect to the original strain. Among the lineages that have been classified as VOCs and spread widely, the recently emerged and rapidly spreading Omicron variant has both the highest number of amino acid mutations and the highest distinctiveness at the polynucleotide level relative to other VOCs. While it is generally expected that any virus will accumulate more mutations over time, it appears that the distinctiveness of VOCs at the polynucleotide level provides additional context beyond the number of mutations acquired. This investigation of the diversity of the nucleotide language explored by SARS-CoV-2 can be considered as a supplement to existing phylogenetic classifications, which themselves were crucial to conduct the presented analyses. Through our publicly available interface (https://academia.nferx.com/GENI), this metric will be continually applied as new variants emerge to understand the degree of distinctiveness for any new lineages with respect to prior or current VOCs.

Implications
The fact that Delta sequences collected during December 2021 are more distinctive with respect to the original strain than Delta sequences collected during April or July 2021 indicates that this metric is at least partially reflective of evolutionary time. This finding is expected given the known evolution of Delta sublineages with additional mutations compared to the original VOC (25). Consistent with this expectation, we have also observed this pattern of increasing distinctiveness of viral variants and VOCs over time, reflecting the natural evolution of the virus. However, the increased distinctiveness of Omicron sequences compared to contemporaneous Delta sequences (i.e. collected in December) interestingly suggests that Omicron may be more distinctive than would be expected from evolutionary time alone. Thus, sequential surges of COVID-19 can be driven by increasingly distinctive VOCs, which continue evolving toward more distinctive lineages during their circulation and are ultimately supplanted by a lineage which is ahead of schedule in terms of genomic distinctiveness. This is consistent with the previous suggestion that VOCs can emerge because of saltational events that occur over extended incubation times in immunocompromised individuals who clear SARS-CoV-2 less efficiently (26–28).

The ability of this metric to predict viral fitness and which variants are most likely to outcompete others needs to be further investigated. Currently, the likelihood that a new lineage will cause surges is speculated based on the number and constellation of mutations that it harbors, particularly based on the Spike protein’s mutations. For example, because the Omicron variant was so heavily mutated as compared to prior strains,
mutations in regions of the Spike protein known to be functionally important (e.g. RBD, furin cleavage site, NTD), it was quickly hypothesized that Omicron would drive surges widely. Given that the propensity of an emerging variant to drive surges is related to several factors including immune evasive capacity (3), we hypothesize that variants which are more distinctive compared to lineages that have previously circulated or are contemporaneously circulating in the same region may be likely candidates for new VOCs. This hypothesis is consistent with the observation that the highly distinctive Omicron variant has already overtaken Delta as the dominant variant in many regions of the world as of January 5, 2022, whereas the highly mutated but less distinctive IHU variant (which emerged before Omicron) was apparently unable to outcompete Delta. It should be emphasized that the IHU variant has several distinctive polynucleotide n-mers that are not observed in previous VOCs which circulated significantly across the globe (Fig. 4C). It will be interesting to determine whether such n-mers that are distinctive to a nondominant but highly mutated variant are “recycled” by VOCs that arise in the future. Additionally, we observe high distinctiveness outside of the Spike protein, specifically for the BA.2 lineage (1 of the Omicron lineages), suggesting that distinctiveness outside of the Spike protein is also an important factor. One benefit of our distinctiveness metric is that even a single sequence of a potential new lineage can be rapidly assessed for distinctiveness relative to prior VOCs or time- and geography-matched SARS-CoV-2 genomes, with further statistical confidence gained as more sequences from such a new lineage are deposited. The free availability of our software (GENI) on the internet should...
Fig. 5. Assessment of the impact of new VOCs on the genomic distinctiveness of prior VOCs. (A) Distribution of the number of distinctive 9-mer polynucleotides contained in reference, Alpha, Beta, and Gamma genomes collected before June 2021. Genomes from other VOCs (including Delta) did not contribute at all to this analysis. (B) Distribution of the number of distinctive 9-mer polynucleotides contained in reference, Alpha, Beta, Gamma, and Delta genomes before June 2021. Only the VOCs shown in the plot (i.e. not Omicron) contributed to this analysis. Abbreviations: VOCs—variants of concern.

Further aid in near-real-time evaluation of such new lineages or sublineages.

Future directions

There are several other questions to which this method can be applied, some of which we have explored preliminarily. For example, we have evaluated genomic distinctiveness between lineages that have previously been or are currently labeled as variants of interest (VOIs) rather than VOCs (3). Here, we do not observe any Delta-like or Omicron-like outliers with respect to their polynucleotide distinctiveness (Figure S3A, Supplementary Material), which is consistent with the fact that these lineages have not been widely responsible for large COVID-19 surges. As a control, we also confirmed that seasonal human-infecting coronavirus (HCoV) genomes, which are similar in size to SARS-CoV-2 genomes, are substantially more distinct from the original (Wuhan) SARS-CoV-2 strain than any VOCs (Figure S3B and C, Supplementary Material). It should be noted that the 2D visualization provided by this analysis (Figure S3B, Supplementary Material) does not imply a transition of VOCs toward a more seasonal coronavirus-like genome, although methods to understand whether such an evolutionary process is likely to occur at some point will be important to develop as part of follow-up research studies. Others have recently reported that infections with Omicron are less likely to cause severe disease than Delta, but these analyses are limited by small sample sizes and confounding variables such as vaccination status (including time since vaccination and booster doses) and higher rates of Omicron infections among young individuals (29–31).

Our results are suggestive of evolutionary pressure as a driver toward genomic distinctiveness, as new variants with high distinctiveness are found to outcompete contemporary variants with lower distinctiveness. Additional analysis, using alignment-based methods to determine positive selection during viral evolution, such as the dN/dS ratio (32,33), SLAC, FEL, REL (34,35), and MEME (36) metrics, could be used in future studies to assess whether the observed increase in distinctiveness coincides with increased positive selection in VOC lineages. Previous studies have shown evidence for positive selection during the early adaptation of the SARS-CoV-2 virus to human hosts (37–39), as well as intrahost purifying selection (40), and negative selection during the pre-VOC era (41). Methodology developed to assess patterns of selection in cancer and somatic tissues (42) could enable analysis of positive selection in sets of closely related sequences under strong positive selection pressure, and with frequent and
functionally relevant deletion and insertion mutations, such as is the case for a comparison between SARS-CoV-2 VOCs. It would be particularly interesting to combine a suitable metric of positive selection with the metric for polynucleotide distinctiveness described in this manuscript, as this might reveal specific genes, or parts of genes, that exhibit high positive selection in highly distinct variants. This could provide further insight into which specific parts of the viral sequence are mutated at greater than expected rates to yield genomically distinct variants.

Methodological considerations and study limitations

Most of the analyses presented here refer to 9-mer polynucleotides, but it is reasonable to consider polynucleotides of various lengths. Indeed, we found that this metric was robust to smaller and larger polynucleotides, with particularly strong separation between some VOCs observed for n-mers between 15 and 30 nucleotides. The presented results based on 9-mers should be considered as an example to illustrate the utility of this metric rather than the definitive resolution. There are also some limitations to these analyses. First, the number of Omicron sequences currently available in the GISAID and NCBI databases is low compared to other VOCs such as Delta. Our protocol, which samples genomes with replacement, could result in oversampling of Omicron sequences. This limitation will be addressed in the coming months as more Omicron sequences are deposited. Additionally, it is important to note that the sequence databases used here contain predominantly sequences that are collected in the United States and Europe (Figure S6, Supplementary Material). This sequencing bias is a major limitation in this and other studies on trends in viral evolution, as regions in which VOCs were first identified are underrepresented. Future initiatives that address the sequencing imbalance will be essential to improve studies of SARS-CoV-2 evolution and for increased vigilance on the emergence of future VOCs. Second, while we consider all sliding nucleotide 9-mers, it is also worth exploring similar metrics of genomic diversity while constraining to protein-coding nucleotide n-mers or amino acid n-mers themselves. Third, both methods presented here compare 1 lineage to 1 or many others, and thus they are sensitive to the lineage composition in the complement

Fig. 6. Genomic distinctiveness increases within the Delta lineage over time. (A) Histogram showing the distribution of distinctive 9-mer counts for all VOCs, with Delta genomes divided temporally into 3 groups: early (collected during April 2021), middle (collected during July 2021), and late (collected during December 2021). The distributions shown for the other variants are identical to those shown in Fig. 1(C), and the Alpha, Beta, and Gamma distributions are shown here in gray (not included in panel legend). Note that each subanalysis of temporally restricted Delta sequences was performed independently such that in any round (experiment), only 1 Delta genome was compared to the genomes of other VOCs. That is, early, middle, and late Delta sequences were not compared to each other during the derivation of their distinctive 9-mers. (B) and (C) Pairwise Cohen’s D and JS divergence values between the distinctive 9-mer distributions shown in (A). Abbreviations: μ—mean, IQR—interquartile range; and VOCs—variants of concern.
For example, several Delta sublineages show relatively low polynucleotide distinctiveness through the $A^* (1-B)$ metric, but this is likely due to the fact that they are being compared to other Delta sublineages which are highly related and quite prevalent in the dataset (Figure S4B, Supplementary Material). Finally, further investigation into the relationship between our genomic distinctiveness metrics and other features such as phylogenetic depth and evolutionary time are warranted, although we have performed preliminary sensitivity analyses which suggest that our metrics provide additional value beyond these features.

### Concluding remarks

In summary, we present here a new methodology and an internet resource that will enable researchers to rapidly assess the distinctiveness of new SARS-CoV-2 lineages relative to any prior or contemporary lineages. We encourage further investigation into whether this method can help to classify lineages as VOCs earlier, study how vaccination impacts SARS-CoV-2 genomic diversity, and determine if or when SARS-CoV-2 will transition toward endemicity or seasonality.

### Methods

#### Quantification of number of distinct n-mers for SARS-CoV-2 VOCs

The number of distinctive n-mers for SARS-CoV-2 sequences from the original reference strain (PANGO lineage A) and 5 prior or current VOCs (Alpha, Beta, Gamma, Delta, and Omicron) were calculated using genomes obtained from the GISAID database (https://www.gisaid.org) (6, 7) and the NCBI database (https://www.ncbi.nlm.nih.gov/sars-cov-2/). Specifically, genomes were sampled with replacement for each VOC by mapping to their corresponding PANGO lineages. We generated 100,000 sets of 6 genomes (using a total of 440,845 unique genomes) and then computed the number of unique 9-mer polynucleotides for every combination of the 6 VOCs in each of the 100,000 sets. For a given genome, 9-mer polynucleotides that contain non-ACGT characters were discarded. We refer to this overall process as performing 100,000 experiments. In each experiment, DN9s for a given variant were defined as those which were present in the genome of that variant but not in the genomes of any other variants (schematic of the methodology provided in Figure S5A (Supplementary Material)). We also determine the distribution of prevalent DN9s across different SARS-CoV-2 genes. Here, DN9s that occur in at least 50,000 experiments (out of 100,000 experiments) are deemed prevalent. This procedure was repeated for n-mer sequences of various lengths, including 3, 6, 9, 12, 15, 18, 21, 24, 30, 45, 60, 75, 120, and 240 nucleotides.

In cases where multiple subsets of sequences for 1 variant of concern were plotted in the same panel (Figs 4B and 6A), we did not consider n-mer overlap between the sublineages in evaluating the number of distinct n-mers. This was done to avoid significant left-shifts in the subset distributions, as there will necessarily be significant overlap between the polynucleotide fragments of related subsets (i.e. BA.1 and BA.2, or Delta-July 2021 and Delta-December 2021). Variant subsets that were evaluated in this manner are plotted using dashed lines for their distributions.

For each pair of variants, we computed the Cohen’s D between the distributions of distinctive n-mer sequence counts using the following equation:

$$\text{Cohen’s } D(X, Y) = \frac{\text{Mean}(X) - \text{Mean}(Y)}{\text{StDev_pooled}},$$

where, $\text{StDev_pooled} = \sqrt{\frac{(N_X - 1)\text{StDev}(X)^2 + (N_Y - 1)\text{StDev}(Y)^2}{N_X + N_Y - 2}}$.

The JS divergence values were computed using the SciPy package (version 1.7.3) in Python (version 3.7.10).

#### Mutational Load of SARS-CoV-2 VOCs

For sequences obtained from GISAID, we determined the mutational load at the protein level, defined as the number of amino-acid mutations (insertions, substitutions, and deletions as
reported in GISAID) relative to the originally deposited Wuhan-Hu-1 strain of SARS-CoV-2 (GenBank accession number MN908947: https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3/) (19). Using the approach as that used to calculate the distinct n-mer distributions, we generated a distribution of the mutational load for each VOC. Specifically, we compared the number of amino-acid mutations per sequence in samples of 6 sequences for each of the variants tested. By repeating this comparison for the 100,000 experiments, we generate a distribution per variant for the mutational load. We then performed pairwise comparisons between all variants by calculating Cohen’s D and JS divergence of the resulting distributions.

Phylogenetic reconstruction of SARS-CoV-2 VOCs
The multiple sequence alignment (MSA) of subsampled SARS-CoV-2 genomes (3,095 genomes randomly sampled from “Open data”) was derived from the Nextstrain database (https://nextstrain.org/), accessed on 20th December, 2021 (43). The MSA was further processed to retain 1,539 sequences belonging to the strains of interest: Alpha (97 genomes), Beta (20 genomes), Gamma (28 genomes), Delta (1,380 genomes), Omicron (2 genomes), and PANGO lineage A (12 sequences; Table S1, Supplementary Material). The pruned MSA was then used to calculate the phylogenetic distance using 4 established metrics: (1) Tajima–Nei (21), (2) Kimura (24), (3) Jukes–Cantor (22), and (4) Tamura (23). The pairwise phylogenetic distance between any 2 VOCs was taken as the mean of the phylogenetic distances between all pairs of sequences belonging to those variants.

Calculation of the alternative n-mer distinctiveness metric $A_*$\((1\rightarrow B)\) that incorporates intralineage conservation scores
To compare the distinctiveness of each SARS-CoV-2 lineage versus all other lineages in GISAID and NCBI, we defined a related metric to capture the distinctiveness of n-mers for a specific SARS-CoV-2 (Figure SSB, Supplementary Material). For a given lineage, \(l\), we calculate the following n-mer distinctiveness metric:

\[
n - \text{mer distinctiveness for a lineage, } l = \sum_{\text{mers}} A(l, n) \ast (1 - B(l, n)),
\]

where \(A(l, n)\) is the fraction of sequences of lineage \(l\) that contain a specific n-mer, \(n\), and \(B(l, n)\) is the fraction of sequences not of lineage \(l\) that contain n-mer \(n\). The sum is over all n-mers that are found for lineage \(l\). Results presented here specifically use 9 as the n-mer size. This analysis uses all nucleotide sequences available in the GISAID and NCBI data that do not contain non-ACGT characters; sequences with non-ACGT characters were removed to avoid any potential bias in the prevalence variable \(A\) for n-mers in sequence regions with high uncertainty.

Data sources
All SARS-CoV-2 sequences were downloaded from GISAID (https://www.gisaid.org/). The 2021–12–13 version of the raw FASTA file and supporting metadata were used. For the analysis of the distribution of the number of distinct n-mers, shown in the main text (Figs 1, 2, 4, and 6), the data was further enhanced by including sequences downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/sars-cov-2/, accessed on 2022–01–27), that were not also present in the GISAID data.

Funding
The authors declare no funding.

Supplementary Material
Supplementary material is available at PNAS Nexus online.

Author Contributions
KM, MN, PL, and VS designed the research; KM, MN, BR, PG, TF, PA, SS, and RS performed the research; KM, MN, BR, PL, PG, PA, TH, and VS analyzed the data; BR contributed analytic tools; and KM, MN, BR, PL, PG, TF, PA, SS, RS, and VS wrote the manuscript.

References
1. COVID-19 Map - Johns Hopkins Coronavirus Resource Center. [last accessed 2021 December 13]. https://coronavirus.jhu.edu/map.html.
2. Rambaut A, et al. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology Nat Microbio. 5:1403–1407.
3. CDC. 2021. SARS-CoV-2 variant classifications and definitions. [last accessed 2021 December 13]. https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html.
4. Colson P, et al. 2021. Emergence in Southern France of a new SARS-CoV-2 variant of probably Cameroonian origin harbouring both substitutions N501Y and E484K in the spike protein. Arch Virol. 18:1–6. DOI: 10.1101/2021.12.24.21268174.
5. Singh D, Yi SV. 2021. On the origin and evolution of SARS-CoV-2. Exp Mol Med. 53:537–547.
6. Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data - from vision to reality. Eurosurveillance. 22:30494.
7. GISAID - initiative. [last accessed 2021 December 13]. https://www.gisaid.org/.
8. Niesen MJM, et al. 2021. COVID-19 vaccines dampen genomic diversity of SARS-CoV-2: unvaccinated patients exhibit more antigenic mutational variance. medRxiv. DOI: 10.1101/2021.07.01.21259833.
9. McCarthy KR, et al. 2021. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. Science. 371:1139–1142.
10. Anand P, Puranik A, Aravamudan M, Venkatakrishnan AJ, Soundararajan V. 2020. SARS-CoV-2 strategically mimics proteolytic activation of human ENaC. Elife. 9:e58603.
11. Peacock TP, et al. 2021. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. Nat Microbiol. 6:899–909.
12. Coutard B, et al. 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res. 176:104742.
13. Harvey WT, et al. 2021. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 19:409–424.
14. Kang L, et al. 2021. A selective sweep in the Spike gene has driven SARS-CoV-2 human adaptation. Cell. 184:4392–4400.e4.
15. Suratekar R, et al. 2022. High diversity in Delta variant across countries revealed via genome-wide analysis of SARS-CoV-2 beyond the Spike protein. Mol Syst Biol. 18:e10673. DOI: 10.1101/2021.09.01.458647.
16. McCallum M, et al. 2021. SARS-CoV-2 immune evasion by the B1.427/B1.429 variant of concern. Science. 373:648–654.
17. Garushyants SK, Rogozin IB, Koonin EV. 2021. Template switching and duplications in SARS-CoV-2 genomes give rise to insertion variants that merit monitoring. Commun Biol. 4:1343. DOI: 10.1038/s42003-021-02858-9.
18. Fauci AS, Lane HC, Redfield RR. 2020. Covid-19 - navigating the uncharted. N Engl J Med. 382:1268–1269.
19. Wu F, et al. 2020. A new coronavirus associated with human respiratory disease in China. Nature. 579:265–269.
20. auspice. [Last accessed 2021 December 13]. https://nextstrain.org/ncov/gisaid/global.
21. Tajima F, Nei M. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol. 1:315–329.
22. Jukes TH, Cantor CR. 1969. Chapter 24 - evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. Cambridge (MA): Academic Press.
23. Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol. 9:678–687.
24. Kimura MA. 1980. Simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.
25. Arora P, et al. 2021. Delta variant (B.1.617.2) sublineages do not show increased neutralization resistance. Cell Mol Immunol. 18:2557–2559.
26. Kemp SA, et al. 2021. SARS-CoV-2 evolution during treatment of chronic infection. Nature. 592:277–282.
27. Van Egeren D, et al. 2021. Controlling long-term SARS-CoV-2 infections can slow viral evolution and reduce the risk of treatment failure. Sci Rep. 11:22630. DOI: 10.1038/s41598-021-02148-8.
28. Corey L, et al. 2021. SARS-CoV-2 variants in patients with immunosuppression. N Engl J Med. 385:562–566.
29. Wolter N, et al. 2021. Early assessment of the clinical severity of the SARS-CoV-2 Omicron variant in South Africa. Lancet. 399:437–446.
30. Madhi SA, et al. 2022. Population Immunity and Covid-19 Severity with Omicron Variant in South Africa. N Engl J Med. DOI: 10.1056/NEJMoa2119658.
31. Maslo C, et al. 2021. Characteristics and outcomes of hospitalized patients in South Africa during the COVID-19 Omicron wave compared with previous waves. JAMA. 327:583–584.
32. Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.
33. Goldman N, Yang ZA. 1994. Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.
34. Pond K, L S, Frost SDW. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 22:1208–1222.
35. Pond SLK, Frost SDW. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 21:2531–2533.
36. Murrell B, et al. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.DOI: 10.1371/journal.pgen.1002764.
37. Berrio A, Gartner V, Wray GA. 2020. Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function. PeerJ. 8:e10234.DOI: 10.7717/peerj.10234.
38. Tang X, et al. 2020. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev. 7:1012–1023.
39. Chaw S-M, et al. 2020. The origin and underlying driving forces of the SARS-CoV-2 outbreak. J Biomed Sci. 27:73.DOI: 10.1186/s12929-020-00665-8.
40. Tonkin-Hill G, et al. 2021. Patterns of within-host genetic diversity in SARS-CoV-2. Elife. 10:e66857. DOI: 10.7554/eLife.66857.
41. Yi K, et al. 2021. Mutational spectrum of SARS-CoV-2 during the global pandemic. Exp Mol Med. 53:1229–1237.
42. Martincorena I, et al. 2017. Universal patterns of selection in cancer and somatic tissues. Cell. 171:1029–1041.e21.
43. Hadfield J, et al. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 34:4121–4123.