PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Unsupervised identification of significant lineages of SARS-CoV-2 through scalable machine learning methods

Roberto Cahuantzi[a,b,1], Katrina A. Lythgoe[c,d,e], Ian Hall[a], Lorenzo Pellis[a], and Thomas House[a]

Edited by Marcus Feldman, Stanford University, Stanford, CA; received October 10, 2023; accepted February 5, 2024

Since its emergence in late 2019, SARS-CoV-2 has diversified into a large number of lineages and caused multiple waves of infection globally. Novel lineages have the potential to spread rapidly and internationally if they have higher intrinsic transmissibility and/or can evade host immune responses, as has been seen with the Alpha, Delta, and Omicron variants of concern. They can also cause increased mortality and morbidity if they have increased virulence, as was seen for Alpha and Delta. Phylogenetic methods provide the "gold standard" for representing the global diversity of SARS-CoV-2 and to identify newly emerging lineages. However, these methods are computationally expensive, struggle when datasets get too large, and require manual curation to designate new lineages. These challenges provide a motivation to develop complementary methods that can incorporate all of the genetic data available without down-sampling to extract meaningful information rapidly and with minimal curation. In this paper, we demonstrate the utility of using algorithmic approaches based on word-statistics to represent whole sequences, bringing speed, scalability, and interpretability to the construction of genetic topologies. While not serving as a substitute for current phylogenetic analyses, the proposed methods can be used as a complementary, and fully automatable, approach to identify and confirm new emerging variants.

machine learning | clustering | phylogenetics | dimensionality reduction | SARS-CoV-2

Significance

New SARS-CoV-2 variants emerged throughout the COVID-19 pandemic that were more infectious, virulent, or immune evasive, emphasizing the importance of studying the unfolding evolution of the virus. Sequencing technology, combined with considerable investment, made large volumes of viral genetic data available, creating a need to find methods which can work with large data while being sensitive to variants' differences. We report on an exploration of word-statistics characterizations, and application of dimensionality reduction and clustering methods, to gain hindsight on roughly 5.7 million sequences, demonstrating that these methods can identify growing lineages in near real time. This work provides methods to rapidly analyse growing volumes of genetic data at low computational cost, and can be considered complementary to phylogenetic approaches.

The rapid spread of SARS-CoV-2 during during the COVID-19 pandemic resulted in major healthcare and societal challenges at the global level. The periodic emergence of new variants that are more transmissible and/or escape host immune responses has given rise to repeated waves of infection that have each produced considerable burdens of disease despite high rates of vaccination and prior infection history (1). There is now extensive effort in identifying worrying new variants at the very earliest stages of their emergence, which at best may enable elimination of these variants before they become established (2), but otherwise enables forward planning that may, in the future, include the timely production of tailored vaccines.

SARS-CoV-2, like other RNA viruses, has a high mutation rate and a short generation time, meaning it evolves extremely rapidly and on the same timescale as transmission. Consequently, phylogenetic analysis has been a powerful approach to monitor the evolution and spread of SARS-CoV-2 (3–5). Most point mutations, or single-nucleotide polymorphisms (SNPs), that appear on phylogenetic trees are neutral or nearly neutral, meaning the mutations themselves have little or no impact on transmission and on whether lineages will grow or die out. However, some mutations do have a selective advantage because they enable more effective transmission in the population at the time of emergence. A hallmark of SARS-CoV-2 evolution has been the emergence and spread of variants of concern (VOCs) and some of their major sublineages, with each having a large collection (or constellation) of lineage-defining mutations, many of which have given these viral lineages transmission advantages (6).

Identifying viral lineages that are likely to be problematic in the future requires considerable effort (3) on tasks including the alignment of sequences to a reference before their incorporation into a phylogenic tree; the designation of new lineages (namely giving them a nomenclature based on their clade location); and the identification of lineages with potentially troublesome mutations or that are expanding quickly (5). The production of phylogenies containing all high-quality SARS-CoV-2 genomes could help in the automation of this process, but with nearly 16 million sequences available in the GISAID database (7) as of July 2023 (and growing), aligning a significant fraction of the sequences and generating a single phylogeny is only possible with extremely large computational resource and by making strong parsimony assumptions (8, 9). Here, we explore less computationally exhausting unsupervised machine learning

Author contributions: R.C., K.A.L., I.H., L.P., and T.H. designed research; R.C. performed research; R.C. and T.H. analyzed data; and R.C., K.A.L., I.H., L.P., and T.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: roberto.cahuantzi@manchester.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2317284121/-/DCSupplemental.

Published March 13, 2024.

methods, that can cope with massive amounts of genetic data, and could be used to identify growing meaningful lineages.

Alignment-free techniques to characterize sequences can be classified into two types: information-theory and word-statistics based (10–12). Information-theoretic approaches rely on the notion of entropy or uncertainty between sequences, whereas word-statistic approaches measure the probabilistic properties of "words" (i.e., substrings) within a single sequence. An advantage of word-statistics methods, which have been used previously for interspecies genetic analyses (13–19), is the representation of each sequence as a vector of n numbers independent from other sequences. This means that distances between sequences can be calculated from their position in n-dimensional space $\mathbb{R}^n$ rather than pairwise.

One of the most widely used word-statistics is the k-mer count (*kmc*, also known as k-words count or bag-of-words), which is the count of the occurrence of words of length k in a sequence formed by all possible nucleotide combinations; a more detailed definition can be found in the Materials and methods section. Another word-statistics method, known as natural vectors (*nv*), retains information on the distribution of words by creating an array of summary statistics values such as count, mean distances to the first site of the sequence, and the variance of these distances. This method inspired us to explore other forms to characterize genetic sequences (20–24). All the previously mentioned characterization techniques can represent a whole genetic sequence using a smaller vector. However, after a systematic comparison, we decided to focus on *kmc*, since it presented the best evaluation according to our chosen clustering comparison metrics (see *SI Appendix*, where the analyses of the rest of the characterization techniques can also be found).

The *kmc* extracted from genetic data can then be processed using unsupervised learning algorithms, which are defined as algorithms that are not reliant on provision of previously known outputs to compute the rules that govern the relationships among input data (25). Dimensionality reduction of the *kmc* output is used to draw out the relationships among the sequence data of sampled viruses by projecting onto lower dimensions. Unsupervised clustering methods can then be applied to this dimensionality reduction projection to identify the structure that emerges within the dataset by grouping similar sequences together. Classical phylogenetic analysis also involves a chain of feature extraction, dimensionality reduction, and unsupervised (agglomerative) clustering, although in a more computationally expensive manner limiting the number of sequences they can cope with.

The dimensionality reduction projection was produced using PaCMAP (26) and cluster detection using CLASSIX (27) and HDBSCAN (28). These methods lack the representation of ancestral relationships present in a phylogeny, but scale in a computationally efficient manner with data volume, so they may prove to be important complementary tools to track the evolution of SARS-CoV-2 as well as other rapidly evolving pathogens.

Results

As detailed in the Materials and Methods section, we worked with 5.7 million high-coverage whole-genome SARS-CoV-2 sequences available in GISAID (7) as up to 19/01/2023. To demonstrate the potential of using alignment-free word-statistics features to represent genetic sequences to unravel the genetic structure of SARS-CoV-2 populations, we first compared a projection of *3mc* (*kmc* with $k = 3$) from unaligned sequences with a maximum parsimony phylogeny generated using IQTree (29) using an alignment of the same sequences shown in Fig. 1. Because of the computational costs of generating phylogenetic trees, we used a stratified subsample of 5,000 high coverage sequences. The multi-dimensional *3mc* features were projected onto a 2-dimensional space using the nonlinear dimensionality reduction tool PaCMAP (26). The points are colored based on the Scorpio labeling by the Pangolin tool (*Materials and Methods*) (4). The similarity of the clusters of the significant VOCs, such as Alpha (B.1.1.7), Delta (B.1.617.2), Omicron (BA.1), and Gamma (P.1), from the 2d PaCMAP projection to the maximum parsimony tree demonstrates the potential of the combination of word-statistics characterization and dimensionality reduction tools to produce interpretable structures for the distribution of meaningful variants in a genetic space.

To measure the quality of the clusters formed by a 3-dimensional PaCMAP projection of the 5.7 million sample data quantitatively, we implemented two clustering algorithms, HDBSCAN (30) and CLASSIX (27) using a set of parameters we denote by *GISAID1*. The Scorpio labeling, produced by Pangolin, was used as a "ground truth" to calculate the cluster similarity metrics adjusted Rand index (ARI) (31) and adjusted mutual information index (AMI) (32) of the clustering matrices. In both metrics, a value of 1 would mean the detected clusters are completely equivalent to those of the "ground truth" (see *Materials and Methods* for further details). It has been argued (33) that these metrics are optimal for different types of clusters: While, ARI is advised for when the clusters present roughly equal sizes, AMI is more appropriate for when the cluster sizes are highly unbalanced. Based on this consideration, AMI would be more meaningful as a metric for the presented dataset. Here, CLASSIX outperformed the more established HDBSCAN, with ARI and AMI values of 0.474 and 0.605 for CLASSIX, versus 0.471 and 0.594 for HDBSCAN. The results of this analysis of the *3mc* feature for the whole GISAID dataset can be seen in Fig. 2.

We were next interested in how the *3mc*-PaCMAP projection performed when restricted to the major structural protein regions. To accurately identify the different gene regions, we used the aligned sequences. This projection was performed using only the PaCMAP parameters of *GISAID1*, since the aim of this analysis is to compare regions of the genome rather than to identify clusters. The plots within Fig. 3 show that while the S and N regions reduce into many clusters associated with different Scorpio labels, the E and M regions are much more homogeneous. The relative homogeneity observed for E and M may be due in part to their relatively short lengths (227 and 668 nucleotides respectively) compared to S and N (1,259 and 3,821 nucleotides respectively), and also possibly due to being subject to stronger adaptive evolution , in particular the existence of antibodies that bind to S and N (34).

We tested the use of this pipeline to detect the emergence of new variants under the hypothesis that they will form new clusters as they appear. For this, we took a subsample of the sequences collected in England ($n = 982{,}496$), from which subsets were taken forming a temporal cumulative progression with 2-week steps based on their submission date. There were some adjustments made to the parameters of the HDBSCAN clustering algorithm to increase the sensitivity of finding small clusters, while CLASSIX parameters were left as in *GISAID1*. This second parametric set-up was called *GISAID2*. The results can be seen in the *Top* row of Fig. 4, where it is possible to see that HDBSCAN follows Scorpio in number of clusters before overshooting its detection, while CLASSIX stays close, although growing more gradually in early 2021. Further refinement of the PaCMAP parameters resulted in an increase of the Pearson correlation

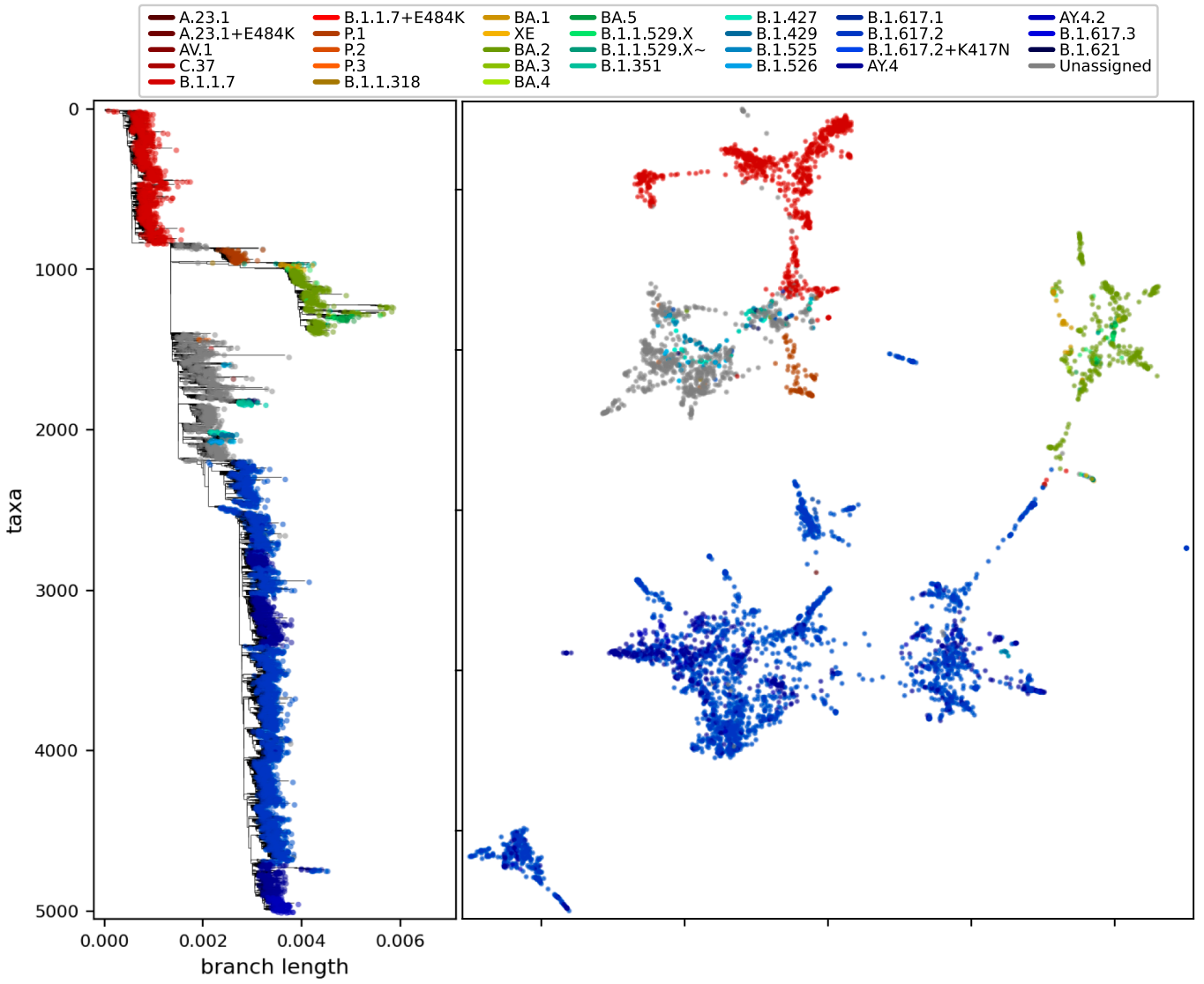

Fig. 1. (*Left*) Comparison of a maximum parsimony phylogenetic tree and (*Right*) *3mc*-PaCMAP projection from a subsampled alignment ($n = 5{,}000$) of high coverage genomes (undefined bases < 1% and length > 29K nucleotides) from those available in GISAID (7), as up to 19/01/2023. The similarities between emerging clusters in both analyses show the potential of the combination of word-statistics features and dimensionality reduction to gain insight into the distribution of variants in a genetic space.

coefficient (r^2) between the CLASSIX cluster detection and the Scorpio labeling, from 0.785 to 0.903. However, this change of parameters reduced the AMI value by a mean of 0.049, from a mean of 0.323 to 0.274, indicating a reduction in the quality of the clusters. This latter parametric set-up was named *England1*.

In terms of computational demands, the entire process from feature extraction to cluster detection can be completed in roughly 30 h on a modern laptop. The *3mc* feature can process a high coverage sequence in roughly 0.12 s, and exploiting parallelization the whole 5.7 million-sequence dataset can be processed in about 14 h. The PaCMAP projection then takes approximately 17 h, while the clustering can be performed in less than 5 min. More details of the process and a diagram of the pipeline can be seen in *SI Appendix*.

To simulate "real-time" cluster growing detection conditions, we processed a series of cumulative subsets to produce "snapshots," namely a *3mc*-PaCMAP projection using the parametric setup *GISAID1* and applying these algorithms to a subset of sequences with submission date earlier or equal to a given date. Once the clusters in any given "snapshot" were identified, the daily counts of the sequences belonging to each cluster were calculated. We applied this to the period during which Delta emerged, revealing dynamics among the clusters similar to the replacement dynamics of Alpha by Delta that were seen within the viral population at the time. Finally, a Gaussian process regression (GPR) as in ref. 35 was applied to smooth the clusters' counts trends, and to calculate the growth rate of clusters of interest (the dominant "cluster 0" and the fast-growing "cluster 15"). This provides further support of the potential of these methods to identify the emergence of meaningful lineages. These results are summarized in the *Bottom* row of Fig. 4. This detection of growing clusters was possible even with a simpler GPR analysis (i.e., one based the normalized proportions for each cluster, unlike the method of ref. 35) for waves of Delta and Omicron, which can be seen in *SI Appendix*.

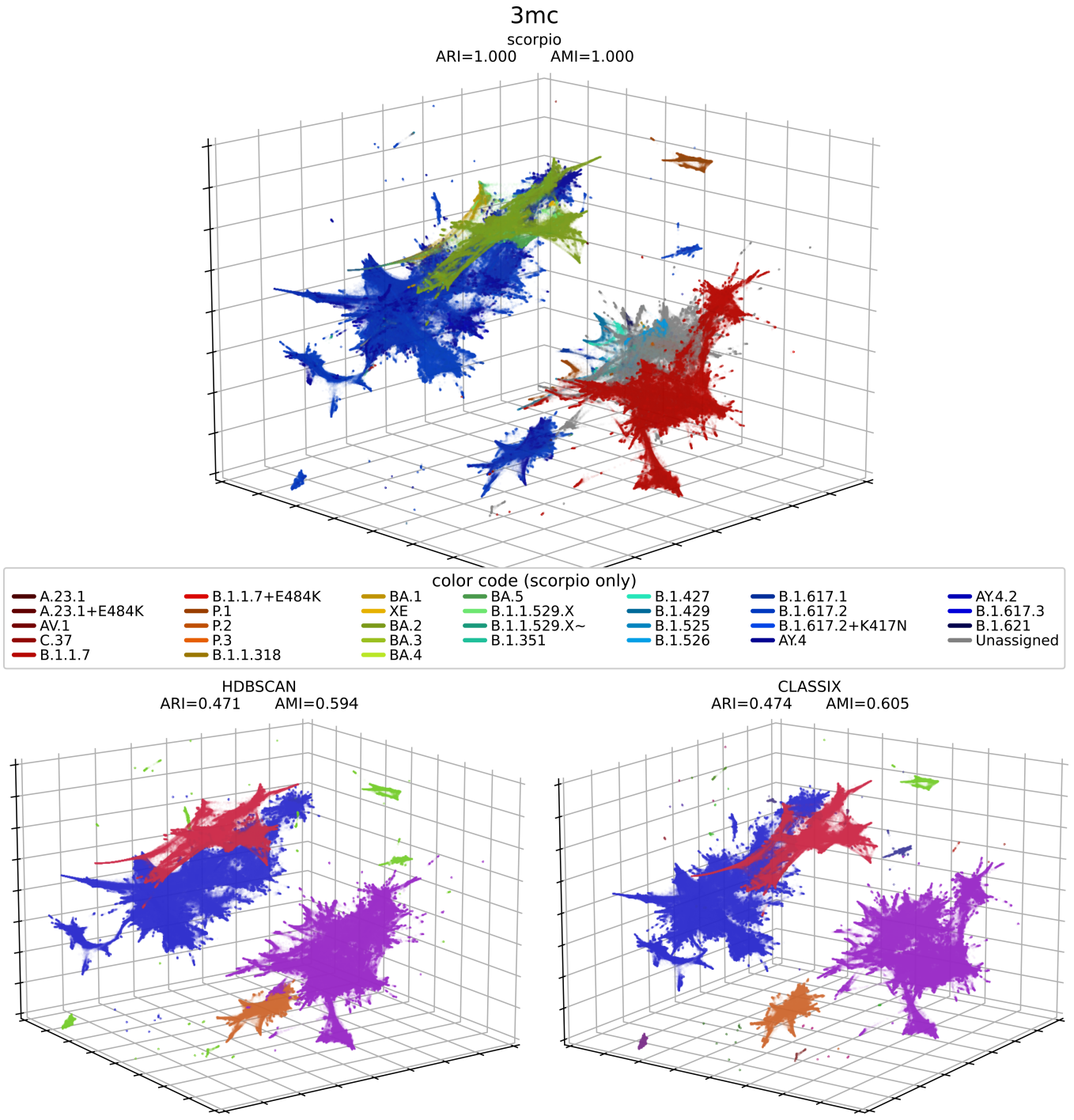

Fig. 2. 3d *3mc*-PaCMAP projection and clustering analysis of genetic topology of 5.7 million sequences. (*Top*, same color code as Fig. 1) Each datapoint correspond to a sequence and is colored by its Scorpio assigned label, (*Bottom-Left*) clustered by HDBSCAN, and (*Bottom-Right*) clustered by CLASSIX algorithm. The projection and clustering was performed using the parametric set-up *GISAID1*, detailed in *Materials and Methods*. In the *Bottom* figures, the color code is arbitrary, showing the unsupervised detection of clusters from the algorithms. The formation of well-defined clusters show the potential of these tools to help us gain insight into the relationships of huge amounts of sequences in a relatively computationally inexpensive fashion. Clicking on the hyperlinks leads to interactive 3d plots.

Discussion

The unprecedented amount of genetic data generated during the pandemic demands the development of methods to analyze it thoroughly, with fluidity and efficiency. Without showing a benefit to curating these data in the future, there is a risk that it will be removed or deleted. The application of characterization for genetic sequences and dimensionality reduction algorithms has the potential to reveal genetic relationships among a huge number of sequences simultaneously. The supervision of these genetic spaces with automated clustering algorithms could provide alternative methods to discover niches of specific viral variants when applied to specific genome regions. Between the two clustering algorithms used, CLASSIX produced results more consistent with existing "gold standard" approaches when compared to HDBSCAN, and with a minor number of parameters, it was

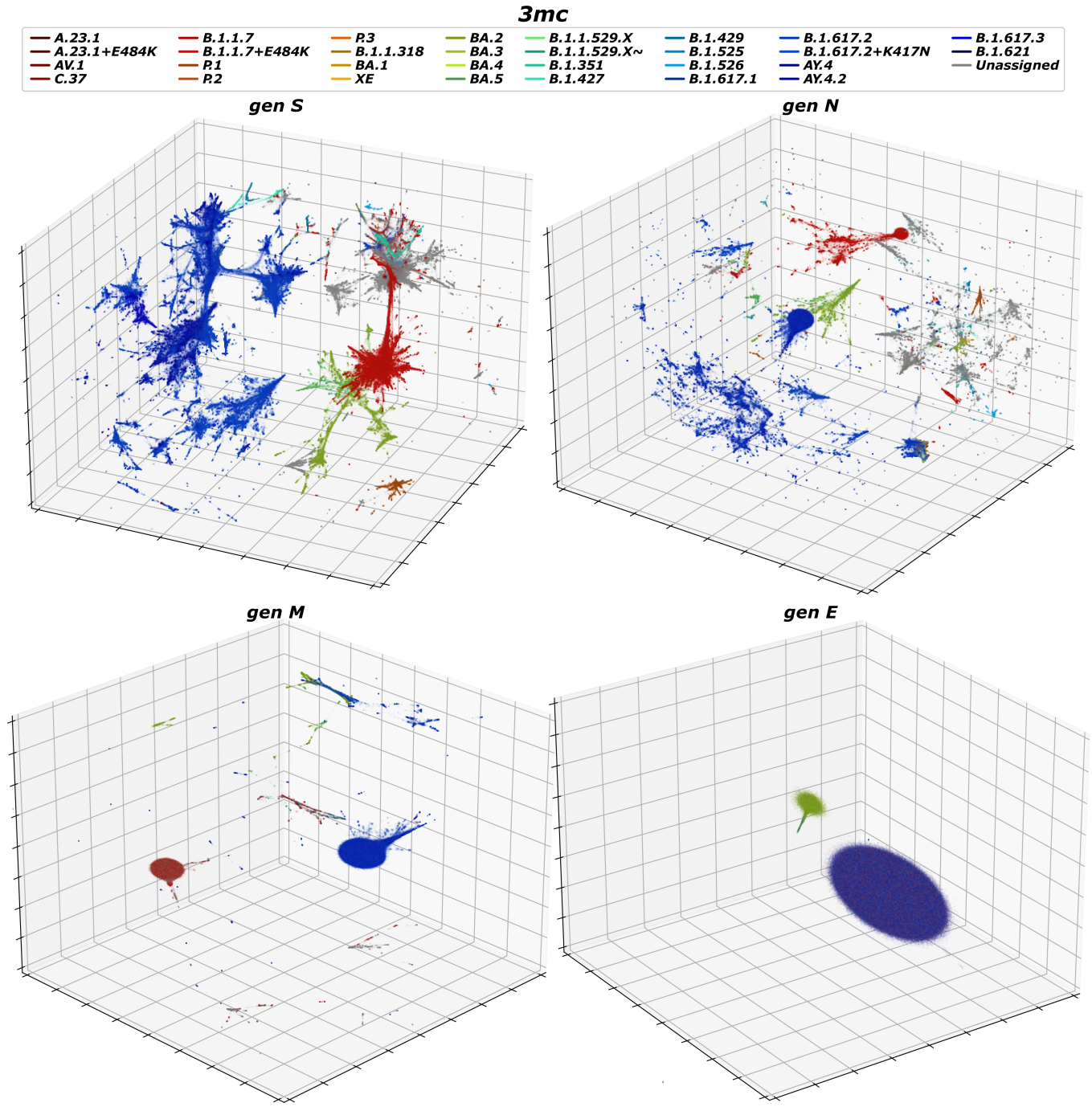

Fig. 3. 3d *3mc*-PaCMAP projection of specific structural protein regions for 5.7 million sequences colored by Scorpio labeling. The formation of clusters allows to see the contribution of the different proteins to the differentiation within the viral population. While proteins S and N seem scattered, arguably due to selection pressures, M and E seem to have found stable configurations. For the projections of genes M and E all variants are contained mainly within the two largest clusters, although due to the high volume and overlap among them, it is difficult to distinguish each of the lineages separately giving the impression of being colored outside of the initial color coding. Clicking on the hyperlinks leads to interactive 3d plots.

easier to optimize. CLASSIX was also ten fold faster and in general yielded higher metrics of cluster similarity (AMI and ARI), compared to Scorpio, making CLASSIX a better tool than HDBSCAN for the scope of this project. Nonetheless, more research on the parametric space exploration of HDBSCAN is needed, together with the investigation of other clustering algorithms.

Our analysis serves as a proof of concept, demonstrating the potential use of machine learning methods as an alert tool for the early discovery of emerging major variants, similarly to the early warning signals from TFP-Scanner (36), and other phylogenetic methods under development, but without relying on the need to generate phylogenies. While phylogenetics remains the "gold standard" for understanding the ancestral relationships of viral populations, the machine learning methods we have proposed have the advantage of being able to manage several orders of magnitude more sequences than the current phylogenetic methods and at a low computational cost. In the analysis presented here, the processing of 5.7 million high-coverage sequences was done in 1 to 2 d on a standard modern laptop.

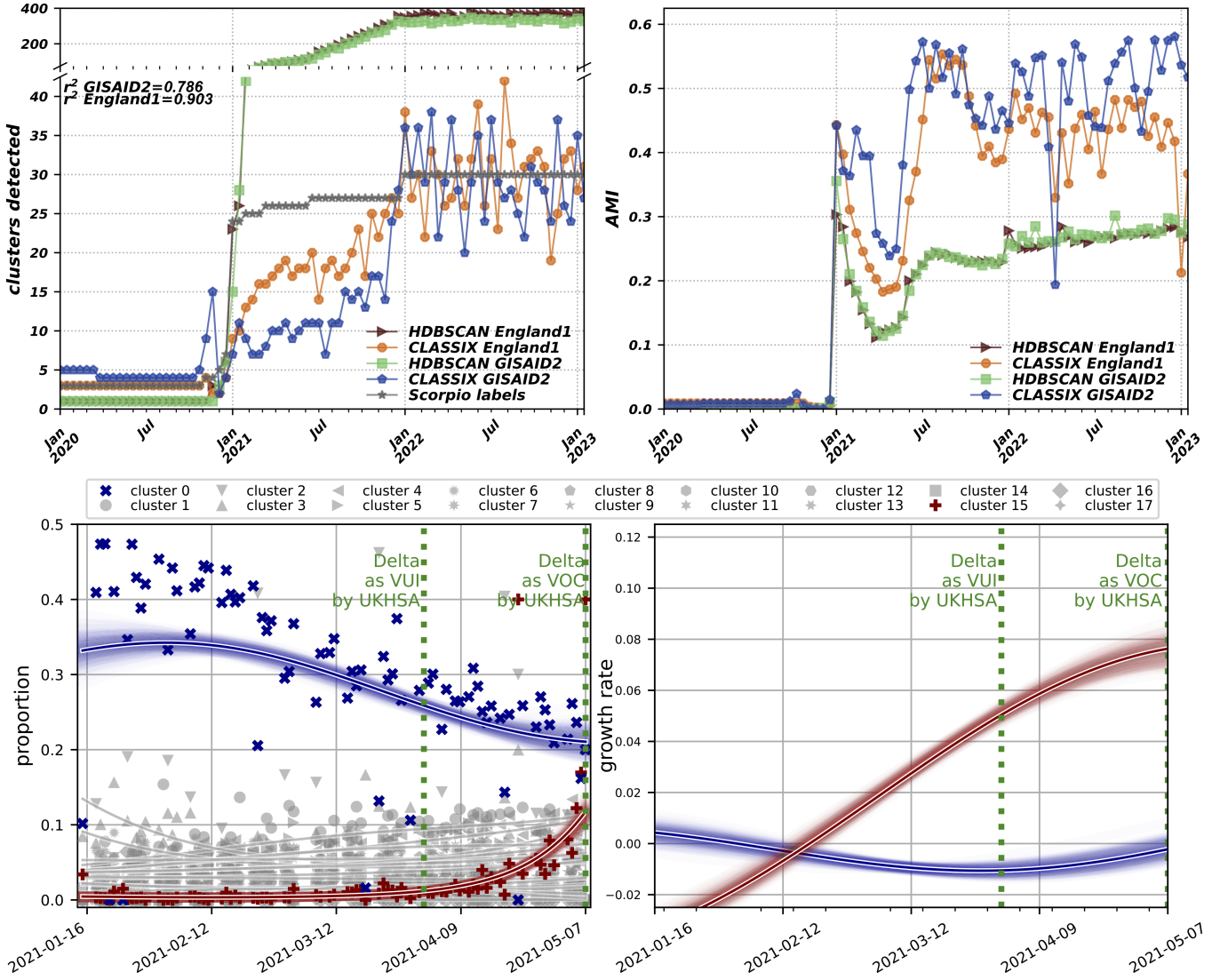

Fig. 4. (*Top-Left*) Number of lineages/clusters detected in the subset of GISAID sequences reported from England. The dimensionality reduction and clustering algorithms were applied to a cumulative temporal progression with resolution of 2-wk and with set-up parameters *GISAID2* and *England1*. (*Top-Right*) For CLASSIX, the contrast between these set-up parameters showed a trade-off, such that while the r^2 to Scorpio is increased from 0.786 to 0.903, the AMI is reduced from a mean of 0.323 to one of 0.274, which implies a reduction of the cluster quality. Furthermore, no major improvement was seen in HDBSCAN clustering. These results point to the potential of *3mc*-PaCMAP projection and clustering detection to identify the possible emergence of new variants by detecting the appearance and growth of clusters. (*Bottom-Left*) Cluster dynamics from the *3mc*-PaCMAP projection and CLASSIX clustering of the data available by 07/05/2021, proving the potential for detection of automatic cluster growth. The initially largest cluster is highlighted in blue and the fastest growing cluster at the end is highlighted in red. Vertical dotted green lines show the times at which UKHSA designated the Delta variant a Variant Under Investigation (VUI), and later VOC. (*Bottom-Right*) Estimated growth rate of clusters of interest.

Balancing the trade-offs of a proper information extraction from the sequences, with computing costs, will be the subject of future work, as will be an evaluation on which features might be most significant for a suitable construction of genetic spaces. Value could be added to these analyses by refinig the characterisation methods through combination of different features (ej. *3mc+envk3*, see Supplementary Information), composed characterisation through the extraction of features by each meaningful genomic region (ej. *3mc(S)+3mc(N)+3mc(M)+3mc(E)*) or applying other characterisation methods such as "protein-*nv*" (37), graphical representation (38), and Fourier power spectrum (39). Furthermore, previous methods for phylogenetic reconstruction have used non-Euclidean distances such as Wasserstein (40), Kullback–Leibler (41), Yau–Hausdorff (38), or Structural Similarity Index Measure (42). Thus, applying these to dimensionality reduction algorithms might generate better representations of the genetic landscape. Finally, applying more sophisticated GPR to smooth the cluster growth counts could improve the detection of growth rate signals.

The methods presented in this paper provide the means to accelerate the identification of emerging pathogen VOCs. The extremely rapid pace of algorithm development and their implementation for dimensionality reduction and unsupervised clustering is expected to continue, and as such produce further valuable analyses that can complement more computationally expensive phylogenetic methods.

Materials and Methods

The process described in this work was run using a laptop with a processor of 11th Gen Intel(R) Core(TM) i7-11800H and 2.30 GHz, with a memory RAM of 16 GB. The full GISAID database (7) was downloaded on 19 January 2023

when it contained 14,617,387 SARS-CoV-2 sequences. These sequences were then filtered to include only complete (>29,000 nucleotides), high-coverage (<1% loci identified as N), human-originated sequences resulting in a total of 5,726,839. Each sequence was aligned to the reference sequence hCov-19/Wuhan/WIV04/2019 (43) using the tool MAFFT v7.453 (44) on an Ubuntu Windows subsystem for Linux v20.04.1 LTS and Biopython v1.78 scripts. Then, Pangolin lineages were obtained by running PangoLEARN v1.18 (4) and taking the "Scorpio call" as "ground truth," to compare to the clusters detected from the PaCMAP projections. The first step was a comparison between the phylogenetic methods and the PaCMAP projection (as seen in Fig. 1). To produce this a subsample of $n = 5{,}000$ from the high-coverage sequences dataset sequences, stratified by Scorpio lineage, was aligned and processed to generate maximum parsimony phylogenetic trees using the tools MAFFT and IQTree v2.2.2.6 (29). This size of subsample was determined by the limiting computational cost of the phylogenetic analysis. The same subsample was then projected on two dimensions using *3mc*-PaCMAP to visualize the contrast between the emerging structures from the two techniques.

To characterize the sequences, we wrote code in Python v3.10.0. The only feature we considered in the main analysis was the k-mer count (referred to as *3mc*, given $k = 3$), which was produced by sliding a window of length k through the whole sequence shifting one position at a time, obtaining $n-k+1$ overlapping strings, where n is the length of the sequence. From there, the occurrences of words of length k formed by all combinations of nucleotides were counted, resulting in a dictionary with the frequencies of each k-mer as values. This means that for a sequence AAAAACGT, this algorithm would extract a *3mc*-dictionary: {AAA:3, AAC:1, ACG:1, CGT:1}. For simplicity, and to keep within the scope of scalability of this project, the value of k was set to 3 for all k-dependent algorithms. Several other word-statistics characterization techniques were analyzed, but we focused on *3mc* which had the best clustering performance, but the results of the analyses using the other characterization techniques can be found in *SI Appendix*.

These word-statistics methods to characterize the sequences were then projected onto low dimensions using unsupervised dimensionality reduction algorithms. Within a machine learning context, unsupervised methods are defined as those producing an automatic learning of computational rules describing the relationship rules among the data using only the input data, unlike supervised methods which find these relationship rules using both input and output data. Unsupervised learning methods are often used to improve interpretability by either human or computational supervised analysis (25). Some of them are nonlinear dimensionality reduction methods such as t-SNE (45) and UMAP (46), as well as linear methods such as principal components analysis which we explored. We decided to use the algorithm PaCMAP v0.5.2.1 (26) which produced robust and interpretable results.

Afterward, the low dimensional space was minmax normalized, namely the lowest value was normalized to 0 and the highest to 1 for all three dimensions of the projection. Then, unsupervised clustering was performed using HDBSCAN v0.8.27 (28, 30) and CLASSIX v0.6.5 (27). The parameters of these clustering algorithms were set to their defaults, with the following exceptions. For HDBSCAN, we set min_cluster_size=200_000, since otherwise this defaults to 5, a value that generates an excessive number of small clusters in a dataset of this size. For CLASSIX, we set radius=0.2 (compared to a default of 0.5) and minPts=500 (compared to a default of 0). These changes were made to improve the algorithm's detection sensibility applied to the whole dataset.

The quality of cluster detection was then evaluated through the metrics of Adjusted Rand Index (ARI) and Adjusted Mutual Information Index (AMI). Searching for the maximization of these metrics, the tuning of PaCMAP's algorithmic hyperparameters was carried out on a 5% subset of the data, stratified by Scorpio labeling, before running the process on the full dataset with the optimized parameters. The *3mc*-PaCMAP projection of the 5% subset was repeated on a 3×3 grid for the PaCMAP parameters with combinations of the hyperparameters MN in {0.25, 0.5, 1.0} and FP in {1.0, 2.0, 4.0}, while the rest of parameters were kept at their default values. The optimal values for the *3mc*-PaCMAP projection were 1.0 and 1.0 for MN and FP respectively. We call this parametric set-up for the clustering and PaCMAP algorithms *GISAID1*. The optimal values however differ for each characterization technique, as detailed in *SI Appendix*.

In addition, to test the detection of emerging clusters through time, the *3mc*-PaCMAP projection and clustering detection methods were applied to a series of cumulative subsamples through time. This analysis only accounted for the subset of GISAID sequences reported from England. The time resolution step was defined as 2 wk. The *3mc*-PaCMAP projection was performed with the same parametric set-up as *GISAID1*, with the exception of a change on the sensitivity of the size of the cluster detection of the HDBSCAN algorithm, since we set min_cluster_size=500 for consistency with CLASSIX. This set-up was called *GISAID2* and resulted in an encouraging, although noisy, detection of clusters for CLASSIX that roughly matched the Scorpio labels, but a tendency of HDBSCAN to overshoot the number clusters detected by Scorpio. Further parametric exploration was made to try to match more closely the Scorpio labeling. The new parametric set-up changed the PaCMAP parameters MN and FP to 0.25 and 2.0 respectively, while keeping the parameters of the clustering algorithms as in *GISAID2*, and was called *England1*. Last, to review the capability for cluster emergence detection of these methods, a Gaussian Process regression was performed for each cluster based on their submission and collection dates, similarly to the technique applied in ref. 35, which allowed calculation of the growth rate signal of any given detected cluster.

Interactive 3D *3mc*-PaCMAP projections plots, Python scripts, and a subsample of the dataset are available at: github.com/robcah/dimredcovid19.

Data, Materials, and Software Availability. The code, fasta files and interactive 3D *3mc*-PaCMAP projections plots are available at https://github.com/robcah/dimredcovid19 (47), with persistent identifier (48).

ACKNOWLEDGMENTS. This project was possible thanks to the support of the Joint Universities Pandemic and Epidemiological Research Consortium (JUNIPER – https://maths.org/juniper/), the Engineering and Physical Sciences Research Council, the Wellcome Trust and the Li Ka Shing Foundation. L.P. gratefully acknowledges the Wellcome Trust and Royal Society (grant 202562/Z/16/Z). T.H. is supported by the Royal Society (grant INF\R2\180067). I.H. is supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emergency Preparedness and Response and the National Institute for Health Research Policy Research Programme in Operational Research (OPERA). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England. L.P., T.H., and I.H. are also supported by the JUNIPER modeling consortium (grant MR/V038613/1) and by the Alan Turing Institute for Data Science and Artificial Intelligence under the EPSRC grants EP/N510129/1 and EP/V027468/1. L.P., T.H., I.H., and R.C. also acknowledge funding from the UKRI Impact Acceleration Account (IAA 386). K.A.L. gratefully acknowledges the Wellcome Trust and Royal Society [grant (107652/Z/15/Z)] and the Li Ka Shing Foundation.

Author affiliations: [a]Department of Mathematics, The University of Manchester, Manchester M13 9PL, United Kingdom; [b]United Kingdom Health Security Agency, University of Oxford, Oxford OX3 7LF, United Kingdom; [c]Department of Biology, University of Oxford, Oxford OX1 3SZ, United Kingdom; [d]Big Data Institute, University of Oxford, Oxford OX3 7LF, United Kingdom; and [e]Pandemic Sciences Institute, University of Oxford, Oxford OX3 7LF, United Kingdom

1. A. Tuekprakhon *et al.*, Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. *Cell* **185**, 2422–2433.e13 (2022).
2. H. Tegally *et al.*, Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nat. Med.* **28**, 1785–1790 (2022).
3. A. Rambaut *et al.*, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
4. A. O'Toole *et al.*, Assignment of epidemiological lineages in an emerging pandemic using the Pangolin tool. *Virus Evol.* **7**, veab064 (2021).
5. J. Hadfield *et al.*, Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
6. Y. Cao *et al.*, BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature* **608**, 593–602 (2022).

7. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
8. Y. Turakhia *et al*., Ultrafast sample placement on existing trees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
9. N. De Maio *et al*., Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, evab087 (2021).
10. A. Zielezinski, S. Vinga, J. Almeida, W. M. Karlowski, Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* **18**, 1–17 (2017).
11. A. Zielezinski *et al*., "Benchmarking of alignment-free sequence comparison methods" (Tech. Rep., Cold Spring Harbor Laboratory, 2019).
12. S. Vinga, J. Almeida, Alignment-free sequence comparison–a review. *Bioinformatics* **19**, 513–523 (2003).
13. M. Kaden *et al*., Learning vector quantization as an interpretable classifier for the detection of SARS-CoV-2 types based on their RNA sequences. *Neural Comput. Appl.* **34**, 67–78 (2021).
14. C. Li *et al*., Phylogenetic analysis of DNA sequences based on K-word and rough set theory. *Physica A: Stat. Mech. Appl.* **398**, 162–171 (2014).
15. Y. Wang, K. Tian, S. S. T. Yau, Protein sequence classification using natural vector and convex hull method. *J. Comput. Biol.* **26**, 315–321 (2019).
16. Y. Li, L. He, R. L. He, S. S. T. Yau, Zika and flaviviruses phylogeny based on the alignment-free natural vector method. *DNA Cell Biol.* **36**, 109–116 (2017).
17. E. S. Allman, J. A. Rhodes, S. Sullivant, Statistically consistent k-mer methods for phylogenetic tree reconstruction. *J. Comput. Biol.* **24**, 153–171 (2017).
18. K. Hatje, M. Kollmar, A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Front. Plant Sci.* **3**, 192 (2012).
19. D. Lebatteux, A. M. Remita, A. B. Diallo, Toward an alignment-free method for feature extraction and accurate classification of viral sequences. *J. Comput. Biol.* **26**, 519–535 (2019).
20. J. Wen, R. H. Chan, S. C. Yau, R. L. He, S. S. Yau, K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* **546**, 25–34 (2014).
21. R. Dong, L. He, R. L. He, S. S. T. Yau, A novel approach to clustering genome sequences using inter-nucleotide covariance. *Front. Genet.* **10**, 234 (2019).
22. Y. Li, L. He, R. Lucy He, S. S. T. Yau, A novel fast vector method for genetic sequence comparison. *Sci. Rep.* **7**, 12226 (2017).
23. M. Deng, C. Yu, Q. Liang, R. L. He, S. S. T. Yau, A novel method of characterizing genetic sequences: Genome space with biological distance and applications. *PLoS One* **6**, e17293 (2011).
24. S. Pei, W. Dong, X. Chen, R. L. He, S. S. T. Yau, Fast and accurate genome comparison using genome images: The extended natural vector method. *Mol. Phylogenet. Evol.* **141**, 106633 (2019).
25. J. Watt, R. Borhani, A. K. Katsaggelos, *Machine Learning Refined: Foundations, Algorithms, and Applications* (Cambridge University Press, 2020).
26. Y. Wang, H. Huang, C. Rudin, Y. Shaposhnik, Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *J. Mach. Learn. Res.* **22**, 1–73 (2021).
27. X. Chen, S. Güttel, Fast and explainable clustering based on sorting. *Pattern Recognit.* 110298 (2024).
28. R. J. G. B. Campello, D. Moulavi, J. Sander, "Density-based clustering based on hierarchical density estimates" in *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, G. Xu, Eds. (Springer, Heidelberg, 2013), pp. 160–172.
29. L. T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2014). https://academic.oup.com/mbe/article-pdf/32/1/268/13171186/msu300.pdf.
30. L. McInnes, J. Healy, "Accelerated hierarchical density based clustering" in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, R. Gottumukkala *et al*., Eds. (IEEE, 2017).
31. W. M. Rand, Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971). 10.1080/01621459.1971.10482356.
32. M. Meilă, Comparing clusterings-an information based distance. *J. Multiv. Anal.* **98**, 873–895 (2007).
33. S. Romano, N. X. Vinh, J. Bailey, K. Verspoor, Adjusting for chance clustering comparison measures. *J. Mach. Learn. Res.* **17**, 4635–4666 (2016).
34. A. S. Heffron *et al*., The landscape of antibody binding in SARS-CoV-2 infection. *PLoS Biol.* **19**, e3001265 (2021).
35. K. A. Lythgoe *et al*., Lineage replacement and evolution captured by 3 years of the United Kingdom Coronavirus (COVID-19) infection survey. *Proc. R. Soc. B: Biol. Sci.* **290**, 20231284 (2023).
36. K. O. Drake *et al*., Phylogenomic early warning signals for SARS-CoV-2 epidemic waves. *eBioMedicine* **100**, 104939 (2024).
37. C. Yu *et al*., Protein space: A natural method for realizing the nature of protein universe. *J. Theor. Biol.* **318**, 197–204 (2013).
38. K. Tian *et al*., Two dimensional Yau-Hausdorff distance with applications on comparison of DNA and protein sequences. *PLoS One* **10**, e0136577 (2015).
39. T. Hoang *et al*., A new method to cluster DNA sequences using Fourier power spectrum. *J. Theor. Biol.* **372**, 135–145 (2015).
40. S. Akhter *et al*., Kullback Leibler divergence in complete bacterial and phage genomes. *PeerJ* **5**, e4026 (2017).
41. H. Zhu, H. Hao, L. Yu, Identifying disease-related microbes based on multi-scale variational graph autoencoder embedding Wasserstein distance. *BMC Biol.* **21**, 294 (2023).
42. I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, L. Vanneschi, Structural similarity index (SSIM) revisited: A data-driven approach. *Exp. Syst. Appl.* **189**, 116087 (2022).
43. P. Okada *et al*., Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from Wuhan to Thailand, January 2020. *Eurosurveillance* **25**, 2000097 (2020).
44. K. Katoh, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
45. L. vd. Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
46. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv [Preprint] (2018). https://arxiv.org/abs/1802.03426 (Accessed 8 October 2021).
47. R. Cahuantzi, DimRedCovid19. GitHub. https://github.com/robcah/dimredcovid19. Deposited 14 November 2023.
48. R. Cahuantzi, robcah/LineageIdentificationByML: Dimension reduction analysis subsample of ∼8k sequences. Zenodo. https://doi.org/10.5281/zenodo.734842. Deposited 4 October 2023.