Genetic cartography reveals ancestral relationships of human pathogenic viruses

Sravani Nanduri[1], **John Huddleston**[2], **Allison Black**[2] **& Trevor Bedford**[2*]

***For correspondence:** trevor@bedford.io (TB)

[1]Issaquah High School, Issaquah, WA, USA, [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Abstract

Introduction

Tracking the evolution of human pathogenic viruses in real time enables epidemiologists to respond quickly to emerging epidemics and local outbreaks. Real-time analyses of viral evolution typically rely on phylogenetic methods. These methods can reconstruct the evolutionary history of viral populations from their genome sequences and estimate states of inferred ancestral viruses including their most likely genome sequence, time of circulation, and geographic location (gen epi papers). Importantly, these methods assume that all sequence data share an evolutionary history represented by the clonal replication of genomes. In practice, the evolutionary histories of many human pathogenic viruses including seasonal influenza viruses, Zika virus, and coronaviruses violate this assumption through processes of reassortment or recombination. Researchers have attempted to compensate for these evolutionary mechanisms by limiting their analyses to specific genes (citation?), concatenating multiple genes despite their different evolutionary histories (citation?), or developing more sophisticated models to represent the joint likelihoods of multiple co-evolving lineages represented by networks rather than trees (Muller). However, several key questions in genomic epidemiology do not require full phylogenetic inference of ancestral relationships and states. For example, genomic epidemiologists commonly need to 1) identify clusters of closely-related genomes that represent regional outbreaks or new variants of concern (Black et al.? MicrobeTrace?), 2) rapidly place newly sequenced viral genomes in the evolutionary context of other circulating strains (USHER, NextClade), and 3) flag low-quality or mislabeled genome sequences for exclusion from their analyses. These common use cases can all be addressed by standard statistical methods including clustering, classification, and outlier detection. These methods make few assumptions about the input data and therefore should be applicable to genomic data that violate phylogenetic assumptions.

To apply these methods to a population of viral genomes, we need metrics to compare genome sequences to each other and algorithms to reduce the highly multidimensional input data ($M \times N$ values for M genomes of length N) to one or two dimensions where clustering, classification, and outlier detection are more tractable. The number of mismatches between any pair of aligned genome sequences, also known as the Hamming distance, provides a natural distance metric for viral genomes. Indeed, most phylogenetic methods start by building a matrix of Hamming distances between all sequences in a given multiple sequence alignment. Many dimensionality reduction al-

gorithms including multidimensional scaling (MDS) (***Hout et al., 2012***), t-SNE (***Maaten and Hinton, 2008***), and UMAP (***McInnes et al., 2018***) accept such distance matrices as an input and produce a corresponding lower-dimensional representation or "embedding" of those data. Alternately, principal components analysis (PCA) only requires the input data to be transformed to a matrix of integers before it can embed those data into a few orthogonal dimensions.

Each of these embedding methods has been applied to genomic data to visualize relationships between individuals and identify clusters of related genomes. Although PCA is a generic linear algebra algorithm that optimizes for an orthogonal embedding of the data, the principal components from single nucleotide polymorphisms (SNPs) represent mean coalescent times and therefore recapitulate broad phylogenetic relationships (***McVean, 2009***). PCA has been applied to SNPs of human genomes (***Novembre et al., 2008***; ***Alexander et al., 2009***; ***McVean, 2009***; ***Auton et al., 2015***) and to multiple sequence alignments of viral genomes (***Metsky et al., 2017***). MDS attempts to embed input data into a lower-dimensional representation such that each pair of data points are as far apart in the embedding as they are in the original data. MDS has been applied to multiple gene segments of seasonal influenza viruses to understand evolutionary relationships between segments (***Rambaut et al., 2008***). t-SNE and UMAP build on manifold learning methods like MDS to find low-dimensional embeddings of data that place similar points close together and dissimilar points far apart (***Kobak and Linderman, 2021***). Both t-SNE and UMAP have been applied to SNPs from human genomes (***Diaz-Papkovich et al., 2019***) and single-cell transcriptomes (***Becht et al., 2018***; ***Kobak and Berens, 2019***).

Although these embedding methods are commonly used for qualitative studies of evolutionary relationships, few studies have attempted to quantify patterns observed in these embeddings and no studies have investigated the value of applying these methods to human pathogenic viruses. To this end, we applied PCA, MDS, t-SNE, and UMAP to genomes from recent populations of seasonal influenza virus A/H3N2, Zika virus, MERS-CoV, and SARS-CoV-2. Each of these viruses have impacted human populations globally in the last decade and have been studied in real time by genomic epidemiologists. For each virus and embedding method, we quantified the relationship between pairwise sequence and embedding distances, identified clusters of closely-related genomes in embedding space, and evaluated the accuracy of clusters compared to expert-defined phylogenetic clades. Finally, we tested the practical application of these methods to identify reassortment and outliers in seasonal influenza viruses. These results inform our recommendations for future applications of these methods including which methods are most effective for specific problems in genomic epidemiology and which parameters researchers should use for each method.

Results

Embedding clusters recapitulate phylogenetic clades for seasonal influenza A/H3N2

Seasonal influenza A/H3N2's hemagglutinin (HA) sequences provide an ideal positive control to test dimensionality reduction methods and clustering. A/H3N2's HA protein evolves rapidly, accumulating amino acid mutations that enable escape from adaptive immunity in human populations (**?**). These mutations produce distinct phylogenetic clades that represent potentially different antigenic phenotypes. The World Health Organization (WHO) Global Influenza Surveillance and Response System (GISRS) regularly sequences genomes of circulating influenza lineages (**?**) and submits these sequences to public INSDC databases like NCBI's GenBank (**?**). These factors, coupled with HA's relatively short gene size of 1,701 nucleotides, facilitate real-time genomic epidemiology of A/H3N2 (**??**) and rapid analysis by the embedding methods we wanted to evaluate.

We identified [NNN] A/H3N2 HA sequences from NCBI's GenBank database (methods) spanning from January 2016 to January 2020. To evaluate the optimal parameters for each embedding method and avoid overfitting to specific datasets, we partitioned these data into a training dataset from 2016–2018 (N=[NNN] sequences) and a test dataset from 2018–2020 (N=[NNN] sequences).

We first analyzed the training data with Nextstrain's seasonal influenza workflow that creates a multiple sequence alignment, infers a time-resolved phylogenetic tree, and assigns clade labels to each sequence based on our own expert-defined clade annotations (**???**). We applied each embedding method to the multiple sequence alignment, identified clusters in the embeddings with HDBSCAN (**?**), and evaluated the accuracy of cluster classifications compared to known clade annotations. We applied this general approach in an exhaustive grid search to identify the optimal parameters for each combination of embedding method and HDBSCAN (see Methods).

All four embedding methods qualitatively recapitulated clade-level groupings observed in the phylogeny (Figure 1). Strains from the same clade generally grouped tightly together in PCA, t-SNE, and UMAP embeddings. While MDS followed this general pattern, it also produced separate pairs of A3 and A4 clusters that did not correspond to meaningful subclades. [Is MDS picking up on other characteristics of the sequence data like the number of Ns? Or maybe MDS needs more dimensions to represent these data and the current constraint of 2 dimensions produces suboptimal results.] All of the embedding methods clearly delineated larger phylogenetic clades into separate spaces (e.g., A1 and A2) and, with the exception of t-SNE, placed related subclades closer together (e.g., A2 and A2/re or the A1b subclades). The t-SNE embedding placed distantly related pairs of clades like 3c3.A and A2 as close together as closely-related clades like A2 and its subclade A2/re. These results suggest that t-SNE maintains both local and global structure, but that our interpretation of the absolute distance between points in these embeddings cannot be linear.

To quantify the apparent maintenance of local and global structure in these embeddings, we calculated the relationship between pairwise genetic distance of genomes and pairwise Euclidean distance of those genomes in each embedding. All four methods maintained a linear relationship between genetic and Euclidean distances for genomes that differed by no more than ≈20 nucleotides (Figure 2). However, PCA and MDS were the only methods that consistently maintained that linearity as genetic distance increased (Pearson's $R^2 = 0.767 \pm 0.000$ and 0.849 ± 0.000, respectively). In contrast, the relationship between genetic and Euclidean distance was nonlinear in t-SNE (Pearson's $R^2 = 0.393 \pm 0.001$) and UMAP (Pearson's $R^2 = 0.397 \pm 0.000$) embeddings. Genomes that differed by more than ≈20 nucleotides were equally as likely to map close together as far apart in these embeddings.

Next, we measured how well clusters of genomes in a given embedding corresponded to our expert clade annotations. For each embedding described above, we applied hierarchical clustering with HDBSCAN to assign cluster labels to each genome. For each pair of genomes, we tested whether both genomes belonged to the same clade and the same cluster. We calculated the accuracy of cluster labels using the Matthew's correlation coefficient (MCC) of the resulting pairwise tests (***Matthews, 1975***). Since we previously identified the optimal HDBSCAN parameter based on this same accuracy metric and dataset, we anticipated that the cluster accuracy would be relatively high. We counted genomes that HDBSCAN could not assign to a cluster as false negatives in our MCC calculation, but we also used this number of unassigned genomes as an additional metric of cluster quality.

As expected, the clusters for each method generally corresponded to larger phylogenetic clades (Figure 3, Table 1). The t-SNE embedding produced the most accurate classification (MCC = 0.756) with 20 clusters and [NNN] genomes not assigned to a cluster. UMAP also accurately classified genomes (MCC = 0.662) with only five clusters and no unassigned genomes. PCA (MCC = 0.368) and MDS (MCC = 0.476) both performed relatively poorly but for different reasons. PCA combined genomes from divergent phylogenetic clades A1 and A2 into the same larger cluster (cluster 4) but managed to assign clusters to all but [NNN] genomes. In contrast, MDS distinguished between most large clades including 3c3.A, A1, and A2, but it also placed closely-related strains from the same clades in two separate clusters (clusters 0 and 5) and failed to assign clusters to [NNN] genomes. Clusters 0 and 5 correspond to the apparently arbitrary splitting of both clades A3 and

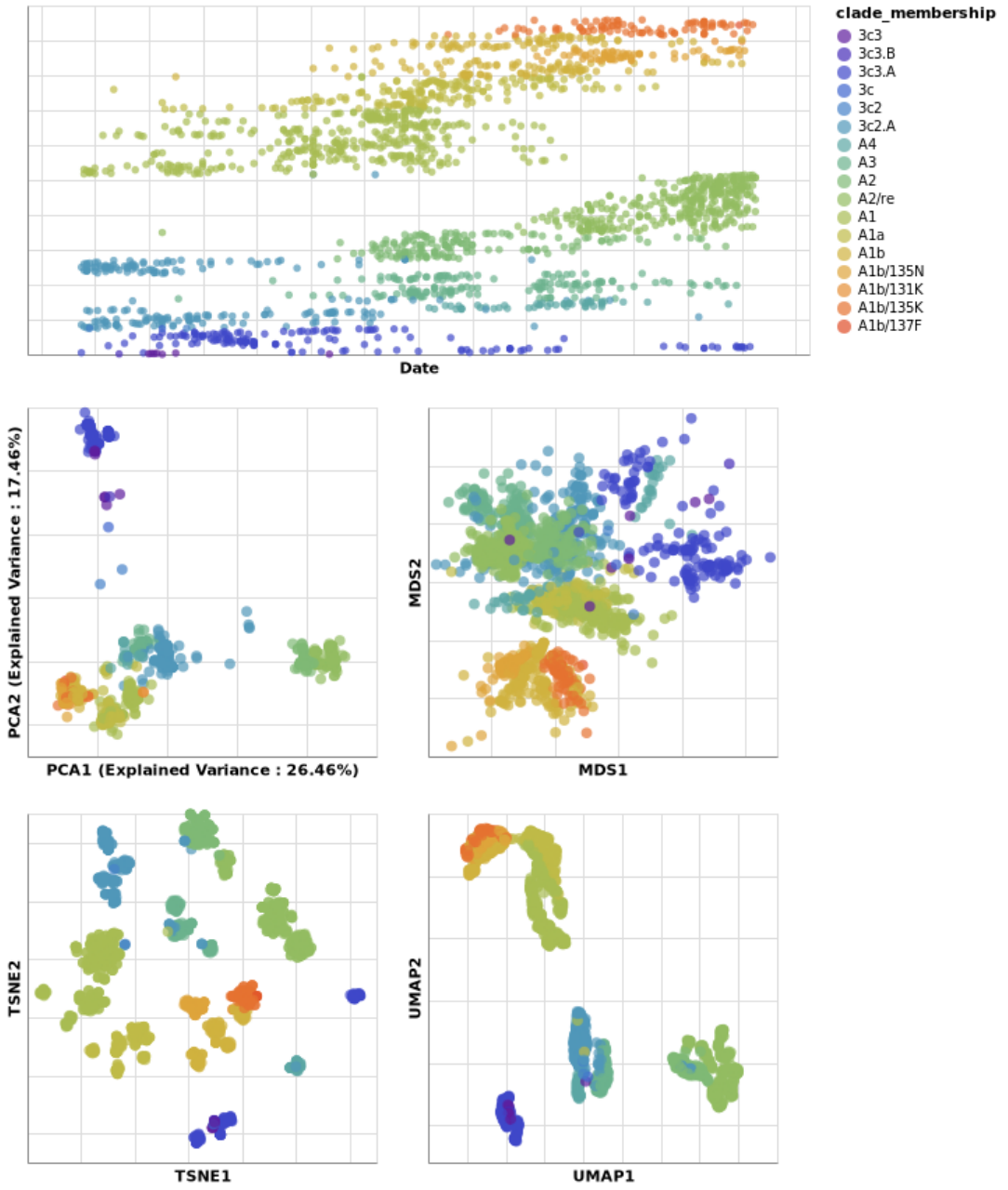

Figure 1. The phylogeny of influenza A/H3N2 viruses (top) shows the evolutionary relationships among viruses including clades, or viruses that share the same mutations and descend from the same common ancestor. Reduced dimensionality embeddings of genetic sequences into two dimensions by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right) generally recapitulate groups of viruses into clades without inferring ancestral relationships. [We should annotate clade membership in the tooltip of the interactive figure.]

A4 into different groups in MDS space described above. These results indicate that nonlinear embeddings of t-SNE and UMAP could be better-suited for clustering and classification than linear embeddings from PCA and MDS.

To understand whether these embedding methods could be used to cluster previously unseen genomes for the same virus, we applied each method to the test dataset spanning 2018–2020, clustered genomes in the embedding space with HDBSCAN, and calculated the accuracy of the

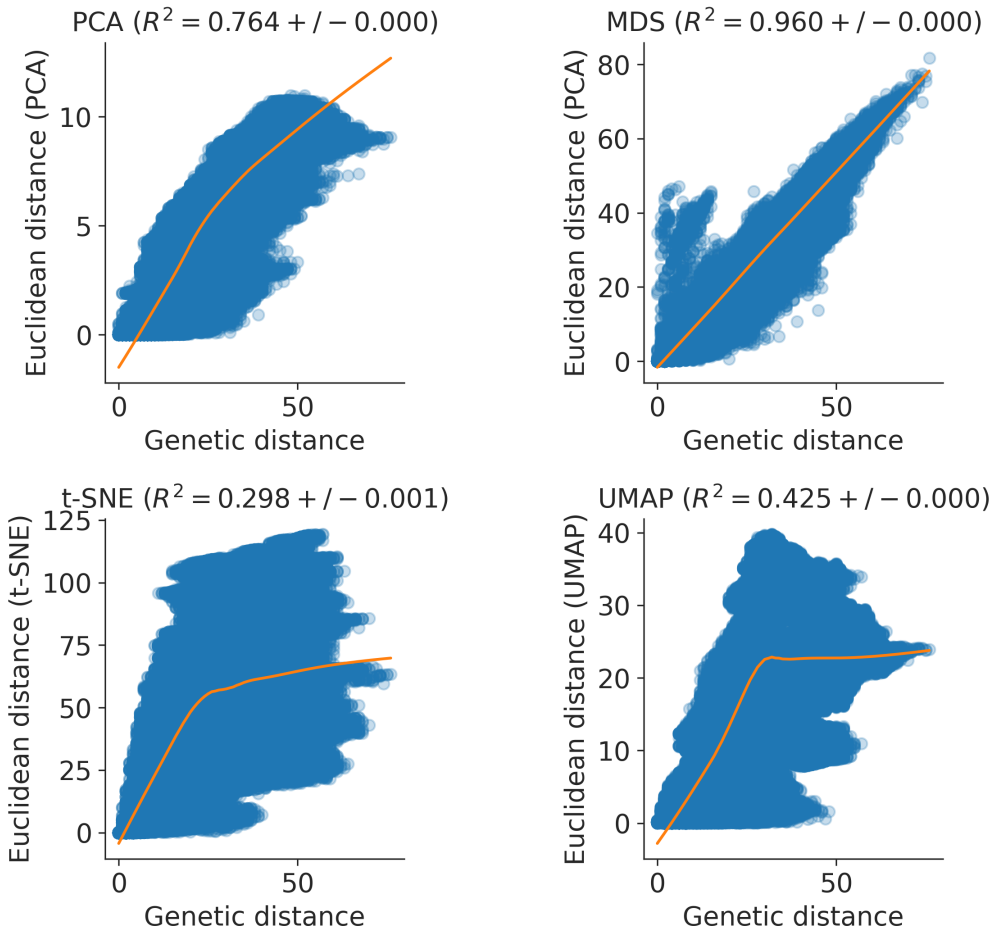

Figure 2. The mapping between Euclidean and Genetic distance assess the strength of both the local and global structure of the embedding recapitulation. The scatterplot for PCA (upper left), MDS (upper right), t-SNE (lower left), and UMAP (lower right) consistently exhibit linear relationships for pairs of strains that differ by around 20 nucleotides.

cluster assignments based on previously defined clades.

Zika virus clusters reveal geographic patterns of the 2013–2018 epidemic

All four dimensionality reduction methods recapitulated phylogenetic patterns observed in the phylogeny (Figure 4). PCA, after imputing missing data, had a similar structure to the findings in Metsky et.al., where the clades were loosely clustered on a continuum of different clades instead of tightly clustered as seen in influenza. Geographical introductions and outbreaks isolated from others were placed at larger Euclidean distances than related introductions. An example is clade c2, an outbreak in Singapore and Thailand separated from the geographical introductions in the Americas. Clade c10 is also a good example of a densely sampled outbreak in Colombia (introduced from Brazil) that forms distinct clusters in all the embeddings. PC1 and PC2 delineate the variance between c2 and the other clades (variance between Asia and the Americas), and PC3 and PC4 are used to show the variance between clade c4 and c3 compared to clade c6 and c9 (variance within the Americas). PC1 and PC2 defined clusters of outbreaks not noted in the phylogenetic tree, such

as a small Brazil-only outbreak as well as a cluster from China and Samoa. Clade c9 is a second parent of an outbreak in Brazil that spread to the US Virgin Islands and Puerto Rico, where c6 is a child outbreak that spread into neighboring countries. All four of the embeddings recognized their relatedness and placed clades c6 and c9 in close proximity to each other. Clade c4, a Central American outbreak that spread to Puerto Rico and other neighboring countries, was not placed closely to clades c6 and c9 even given similar geographical locations and introduction times. This suggests that strains from the same introduction cluster together, and do not cluster just by where they were introduced.

MERS-CoV clusters correspond to host-specific outbreaks

SARS-CoV-2 clusters recapitulate emerging lineage designations

Joint embeddings of hemagglutinin and neuraminidase genomes identify seasonal influenza virus A/H3N2 reassortment events

Embeddings of hemagglutinin genomes identify low-quality or misannotated seasonal influenza viruses

Discussion

Materials and methods

Hyperparameter optimization

To test whether embeddings from each method could recapitulate phylogenetic clades, we performed a grid search of each method's parameter space during which we applied an embedding method to a randomly selected 50% of the sequences in the multiple sequence alignment, identified clusters in the embedding with HDBSCAN (**?**), and calculated the accuracy of the cluster labels for each sequence compared to the known clade labels (see methods). We used the resulting classification accuracies to identify the optimal distance threshold for HDBSCAN. We fixed the HDBSCAN threshold to its optimal value and repeated the same procedure on the other 50% of the sequences. We used the resulting accuracies to identify the optimal t-SNE and UMAP parameters. Finally, we applied each embedding method to the full training dataset with the optimal method parameters, clustered the embeddings with HDBSCAN's optimal distance threshold, and evaluated the accuracy of the cluster classifications.

Data and software availability

The entire workflow for our analyses was implemented with Snakemake (***Mölder et al., 2021***). We have provided all source code, configuration files, and datasets at https://github.com/blab/cartography.

Acknowledgments

Author contributions

SN... JH... AB... TB...

Competing interests

The authors declare that no competing interests exist.

Supplemental Files

References

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009; .

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. A global reference for human genetic variation. Nature. 2015 Oct; 526(7571):68–74.

200 **Becht E**, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction
201 for visualizing single-cell data using UMAP. Nat Biotechnol. 2018 Dec; .

202 **Diaz-Papkovich A**, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure
203 and phenotype heterogeneity in large genomic cohorts. PLOS Genetics. 2019 Nov; https://journals.plos.org/
204 plosgenetics/article?id=10.1371/journal.pgen.1008432.

205 **Hout MC**, Papesh MH, Goldinger SD. Multidimensional scaling. Wiley Online Library. 2012; .

206 **Kobak D**, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019 11; 10(1):5416.

207 **Kobak D**, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP.
208 Nat Biotechnol. 2021 02; 39(2):156–157.

209 **Maaten Lvd**, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008; 9(Nov):2579–
210 2605.

211 **Matthews BW**. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim-
212 ica et biophysica acta. 1975 Oct; https://pubmed.ncbi.nlm.nih.gov/1180967/.

213 **McInnes L**, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.
214 . 2018; .

215 **McVean G**. A genealogical interpretation of principal components analysis. PLoS Genet. 2009 Oct;
216 5(10):e1000686.

217 **Metsky HC**, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, West K, Qu J, Baniecki ML, Gladden-
218 Young A, Lin AE, Tomkins-Tinch CH, Ye SH, Park DJ, Luo CY, Barnes KG, Shah RR, Chak B, Barbosa-Lima G,
219 Delatorre E, et al. Zika virus evolution and spread in the Americas. Nature. 2017 06; 546(7658):411–415.

220 **Mölder F**, Jablonski K, Letcher B, Hall M, Tomkins-Tinch C, Sochat V, Forster J, Lee S, Twardziok S, Kanitz A, Wilm
221 A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. Sustainable data analysis with Snakemake [version 2; peer
222 review: 2 approved]. F1000Research. 2021; 10(33). doi: 10.12688/f1000research.29032.2.

223 **Novembre J**, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al.
224 Genes mirror geography within Europe. Nature. 2008; .

225 **Rambaut A**, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological
226 dynamics of human influenza A virus. Nature. 2008 Apr; https://www.nature.com/articles/nature06945.

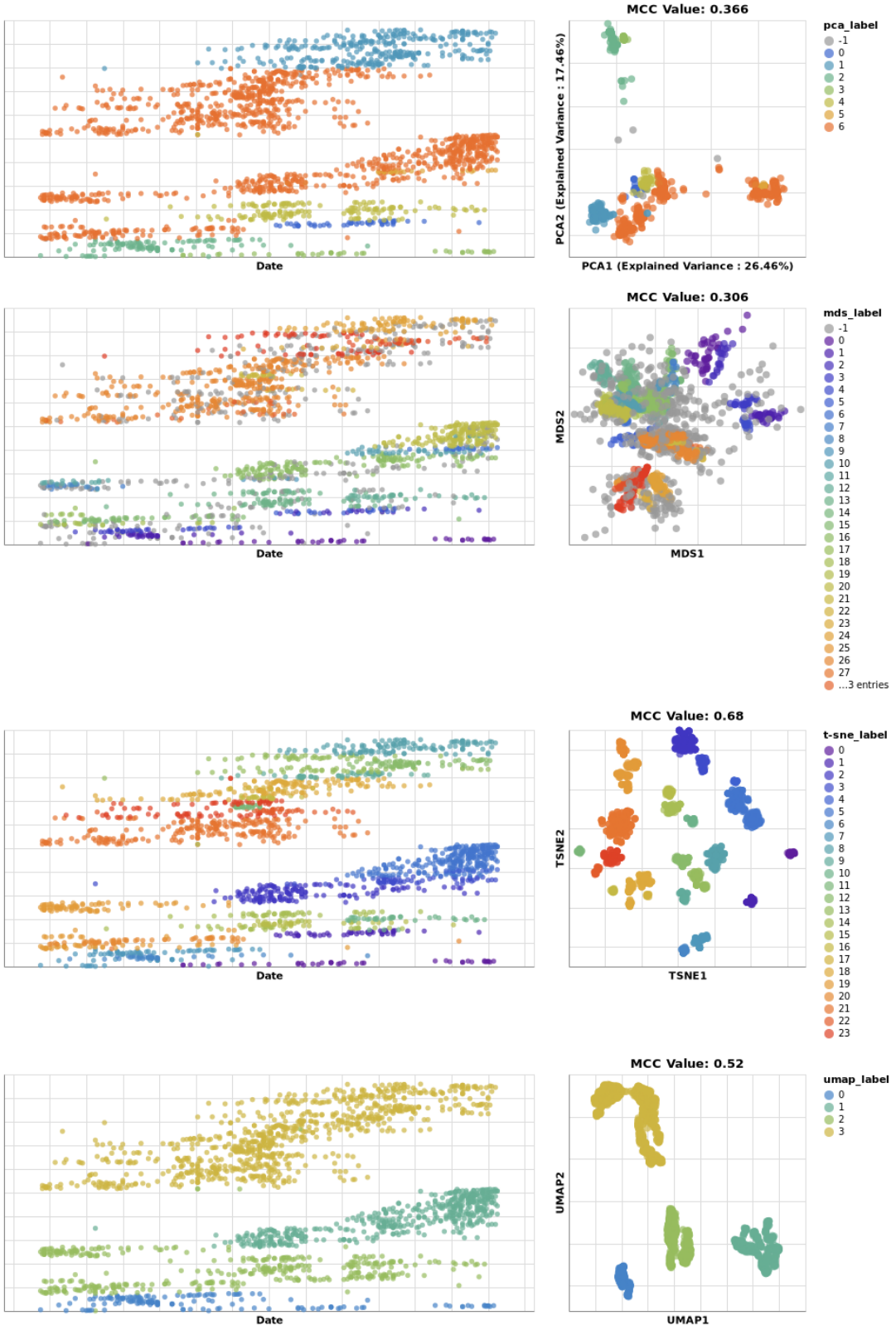

Figure 3. The embeddings colored by their HDBSCAN label, with the distance threshold defined by the threshold that preserved the greatest amount of clade relationships. The chart for PCA (top left), MDS (middle left), t-SNE (middle left), and UMAP (bottom left) generally recapitulate groups of viruses into clades without inferring ancestral relationships, and the trees on the righthand side describes how these clade grouping appear on the tree, which does infer ancestral relations.

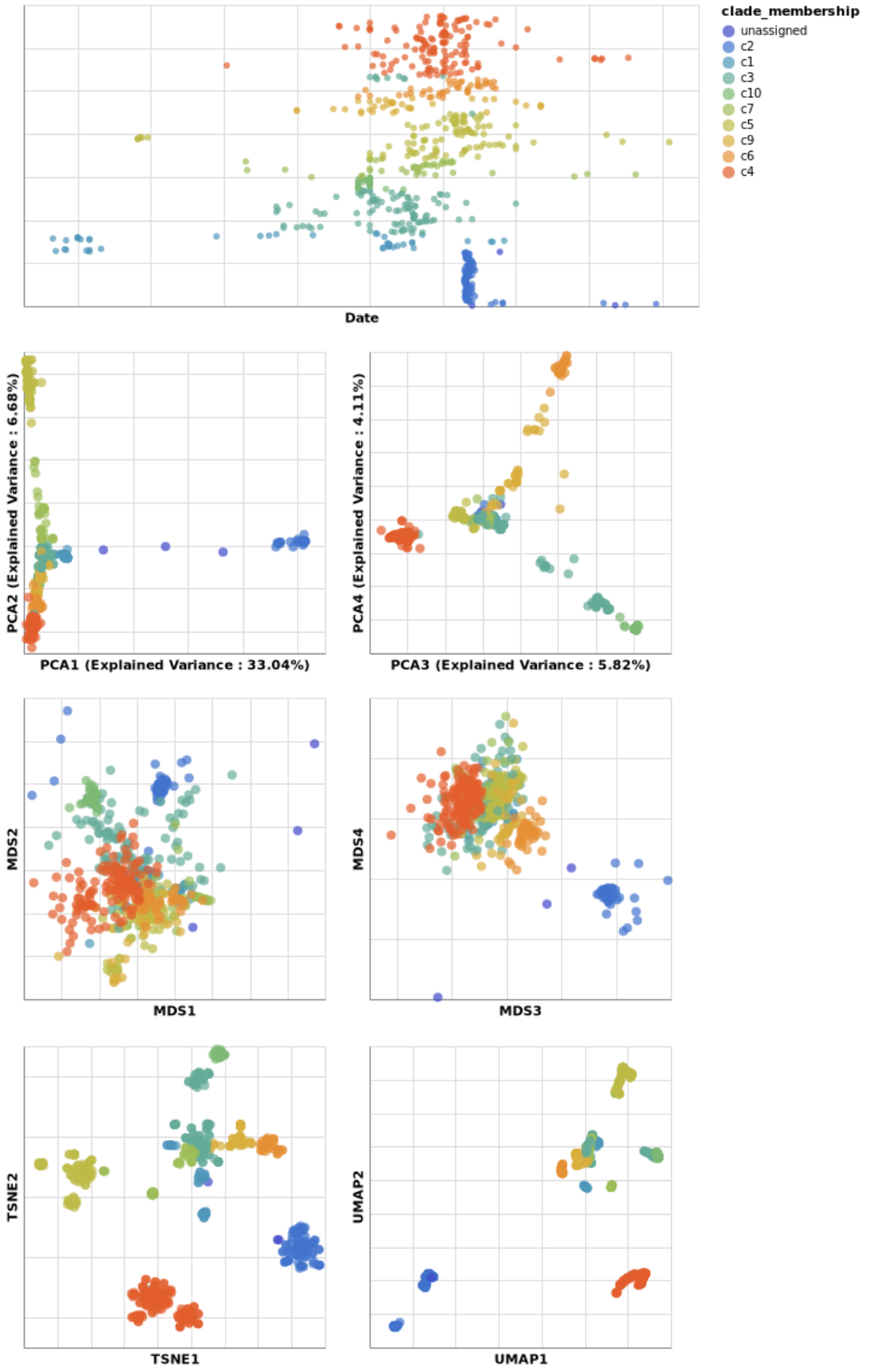

Figure 4. Genetic cartography of Zika strains by dimensionality reduction methods. PCA (components 1 and 2 upper left, components 3 and 4 upper right), MDS (middle left and middle right), t-SNE (lower left), and UMAP (lower right) compared to inferred phylogeny (upper plot). PC1 and PC2 delineate the variance between the Americas and Asia, and PC3 and PC4 are included to better display the variance within the Americas.