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 recombination or reassortment. 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 x N 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-

40 gorithms including multidimensional scaling (MDS) (*Hout et al., 2012*), t-SNE (*Maaten and Hinton,*
41 *2008*), and UMAP (*McInnes et al., 2018*) accept such distance matrices as an input and produce a
42 corresponding lower-dimensional representation or "embedding" of those data. Alternately, prin-
43 cipal components analysis (PCA) only requires the input data to be transformed to a matrix of
44 integers before it can embed those data into a few orthogonal dimensions.

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

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

Results

Embedding clusters recapitulate phylogenetic clades for seasonal influenza A/H3N2
75 As a positive control, we first analyzed hemagglutinin (HA) sequences from the seasonal influenza
76 subtype A/H3N2. A/H3N2's HA protein evolves rapidly, accumulating amino acid mutations that
77 enable escape from adaptive immunity in human populations (**?**). These mutations produce dis-
78 tinct phylogenetic clades that represent potentially different antigenic phenotypes. The World
79 Health Organization (WHO) Global Influenza Surveillance and Response System (GISRS) regularly
80 sequences genomes of circulating influenza lineages (**?**) and submits these sequences to public
81 INSDC databases like NCBI's GenBank (**?**). These factors, coupled with HA's relatively short gene
82 size of 1,701 nucleotides, facilitate real-time genomic epidemiology of A/H3N2 (**??**) and rapid anal-
83 ysis by embedding methods we wanted to evaluate.

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

88 We first analyzed the training data with Nextstrain's seasonal influenza workflow that creates a
89 multiple sequence alignment, infers a time-resolved phylogenetic tree, and assigns clade labels to
90 each sequence based on our own expert-defined clade annotations (**???**). To test whether embed-
91 dings from each method could recapitulate phylogenetic clades, we performed a grid search of
92 each method's parameter space during which we applied an embedding method to a randomly
93 selected 50% of the sequences in the multiple sequence alignment, identified clusters in the em-
94 bedding with HDBSCAN (**?**), and calculated the accuracy of the cluster labels for each sequence
95 compared to the known clade labels (see methods). We used the resulting classification accura-
96 cies to identify the optimal distance threshold for HDBSCAN. We fixed the HDBSCAN threshold to
97 its optimal value and repeated the same procedure on the other 50% of the sequences. We used
98 the resulting accuracies to identify the optimal t-SNE and UMAP parameters. Finally, we applied
99 each embedding method to the full training dataset with the optimal method parameters, clus-
100 tered the embeddings with HDBSCAN's optimal distance threshold, and evaluated the accuracy of
101 the cluster classifications. [This is probably too much detail for the main text and could be clarified.
102 I want to make sure we at least explain that the embeddings in Figure 1 don't represent default
103 method parameters. It's also potentially confusing to introduce HDBSCAN so early, but this is a key
104 component of the analysis so maybe it is fine.]

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

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

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

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.]

137   As expected, the clusters for each method generally corresponded to larger phylogenetic clades
138   (Figure 3, Table 1). The t-SNE embedding produced the most accurate classification (MCC = 0.756)
139   with 20 clusters and [NNN] genomes not assigned to a cluster. UMAP also accurately classified
140   genomes (MCC = 0.662) with only five clusters and no unassigned genomes. PCA (MCC = 0.368)
141   and MDS (MCC = 0.476) both performed relatively poorly but for different reasons. PCA combined
142   genomes from divergent phylogenetic clades A1 and A2 into the same larger cluster (cluster 4)
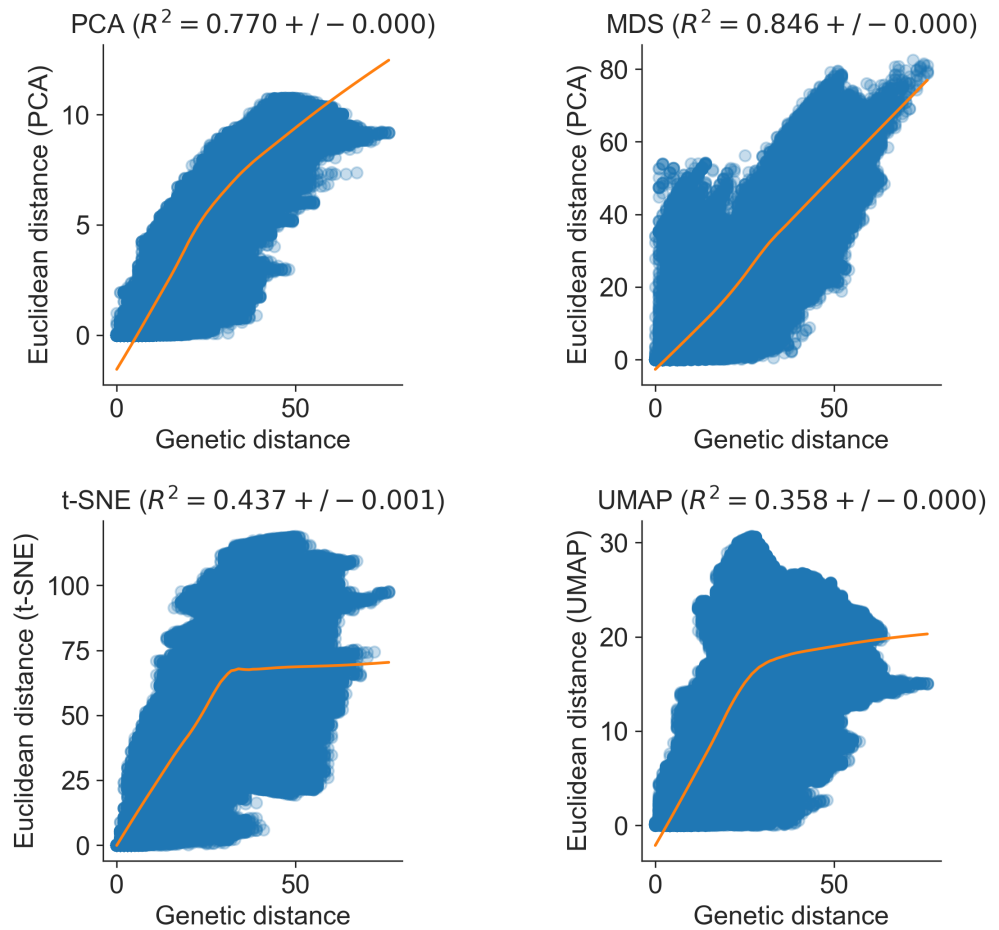
Figure 2. The mapping between Euclidean and Genetic distance assess the strength of both the local and global structure of the embedding recapitulation. The scatterplot for PCA (upper left), MDS (upper right), t-SNE (lower left), and UMAP (lower right) consistently exhibit linear relationships for pairs of strains that differ by around 20 nucleotides.

143 but managed to assign clusters to all but [NNN] genomes. In contrast, MDS distinguished be-
144 tween most large clades including 3c3.A, A1, and A2, but it also placed closely-related strains from
145 the same clades in two separate clusters (clusters 0 and 5) and failed to assign clusters to [NNN]
146 genomes. Clusters 0 and 5 correspond to the apparently arbitrary splitting of both clades A3 and
147 A4 into different groups in MDS space described above. These results indicate that nonlinear em-
148 beddings of t-SNE and UMAP could be better-suited for clustering and classification than linear
149 embeddings from PCA and MDS.

150 To understand whether these embedding methods could be used to cluster previously unseen
151 genomes for the same virus, we applied each method to the test dataset spanning 2018–2020,
152 clustered genomes in the embedding space with HDBSCAN, and calculated the accuracy of the
153 cluster assignments based on previously defined clades.

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

155 All four dimensionality reduction methods recapitulated phylogenetic patterns observed in the
156 phylogeny (Figure 4). PCA, after imputing missing data, had a similar structure to the findings in
157 Metsky et.al., where the clades were loosely clustered on a continuum of different clades instead
158 of tightly clustered as seen in influenza. Geographical introductions and outbreaks isolated from
159 others were placed at larger Euclidean distances than related introductions. An example is clade
160 c2, an outbreak in Singapore and Thailand separated from the geographical introductions in the
161 Americas. Clade c10 is also a good example of a densely sampled outbreak in Colombia (introduced
162 from Brazil) that forms distinct clusters in all the embeddings. PC1 and PC2 delineate the variance
163 between c2 and the other clades (variance between Asia and the Americas), and PC3 and PC4 are
164 used to show the variance between clade c4 and c3 compared to clade c6 and c9 (variance within
165 the Americas). PC1 and PC2 defined clusters of outbreaks not noted in the phylogenetic tree, such
166 as a small Brazil-only outbreak as well as a cluster from China and Samoa. Clade c9 is a second
167 parent of an outbreak in Brazil that spread to the US Virgin Islands and Puerto Rico, where c6 is
168 a child outbreak that spread into neighboring countries. All four of the embeddings recognized
169 their relatedness and placed clades c6 and c9 in close proximity to each other. Clade c4, a Central
170 American outbreak that spread to Puerto Rico and other neighboring countries, was not placed
171 closely to clades c6 and c9 even given similar geographical locations and introduction times. This
172 suggests that strains from the same introduction cluster together, and do not cluster just by where
173 they were introduced.

174 **MERS-CoV clusters correspond to host-specific outbreaks**

175 **SARS-CoV-2 clusters recapitulate emerging lineage designations**

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

178 **Embeddings of hemagglutinin genomes identify low-quality or misannotated sea-**
179 **sonal influenza viruses**

180 **Discussion**

181 **Materials and methods**

182 **Data and software availability**

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

185 **Acknowledgments**

186 **Author contributions**

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

188 **Competing interests**

189 The authors declare that no competing interests exist.

190 **Supplemental Files**

191 **References**

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

194 **Auton A**, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abeca-
195 sis GR, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG,
196 Donnelly P, Eichler EE, et al. A global reference for human genetic variation. Nature. 2015 Oct; 526(7571):68–
**197** 74.

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

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

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

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

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

207 **Maaten Lvd**, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008; 9(Nov):2579–
208   2605.

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

211 **McInnes L**, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.
212   . 2018; .

213 **McVean G**. A genealogical interpretation of principal components analysis. PLoS Genet. 2009 Oct;
214   5(10):e1000686.

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

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

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

223 **Rambaut A**, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological
224   dynamics of human influenza A virus. Nature. 2008 Apr; https://www.nature.com/articles/nature06945.
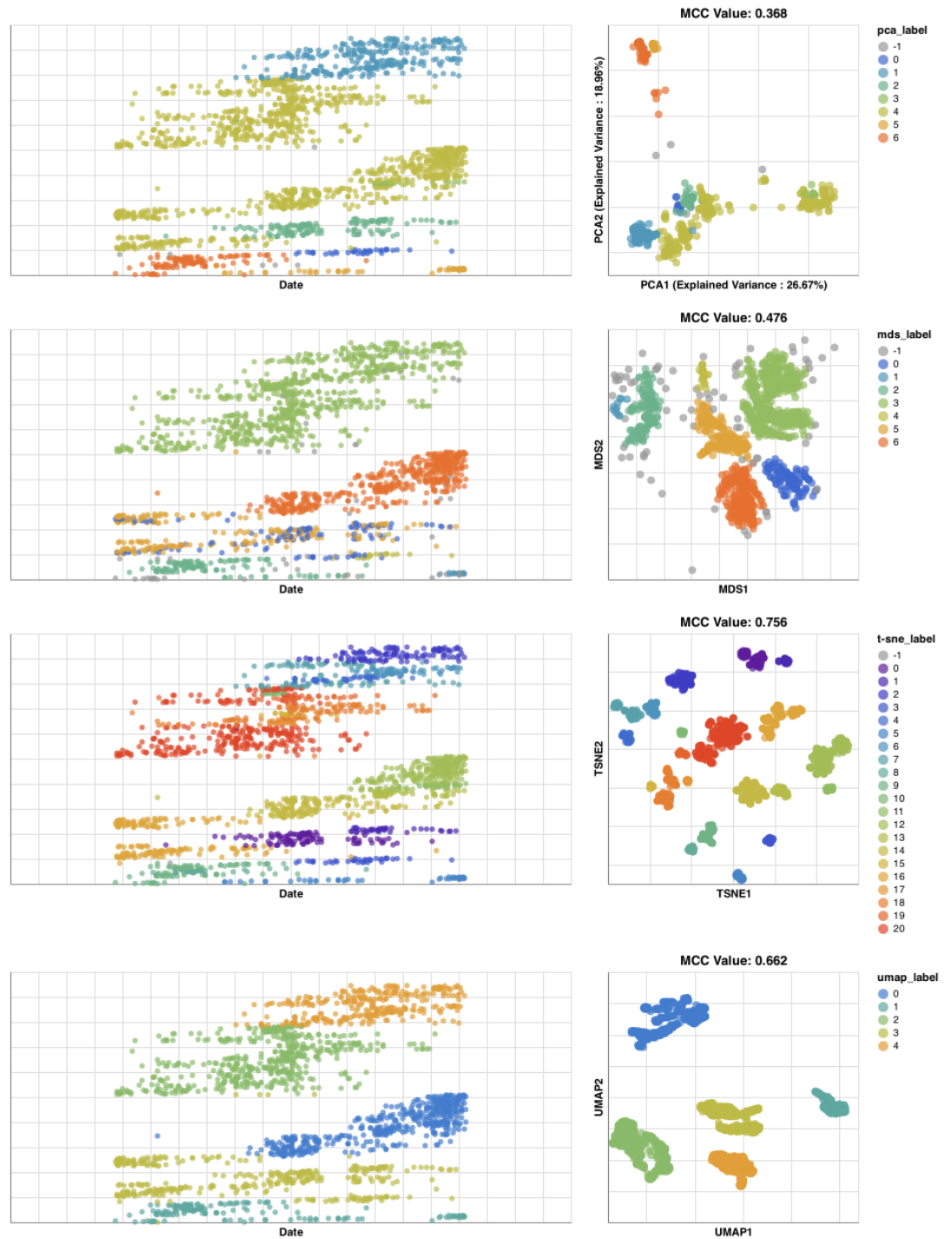
Figure 3. The embeddings colored by their HDBSCAN label, with the distance threshold defined by the threshold that preserved the greatest amount of clade relationships. The chart for PCA (top left), MDS (middle left), t-SNE (middle left), and UMAP (bottom left) generally recapitulate groups of viruses 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.
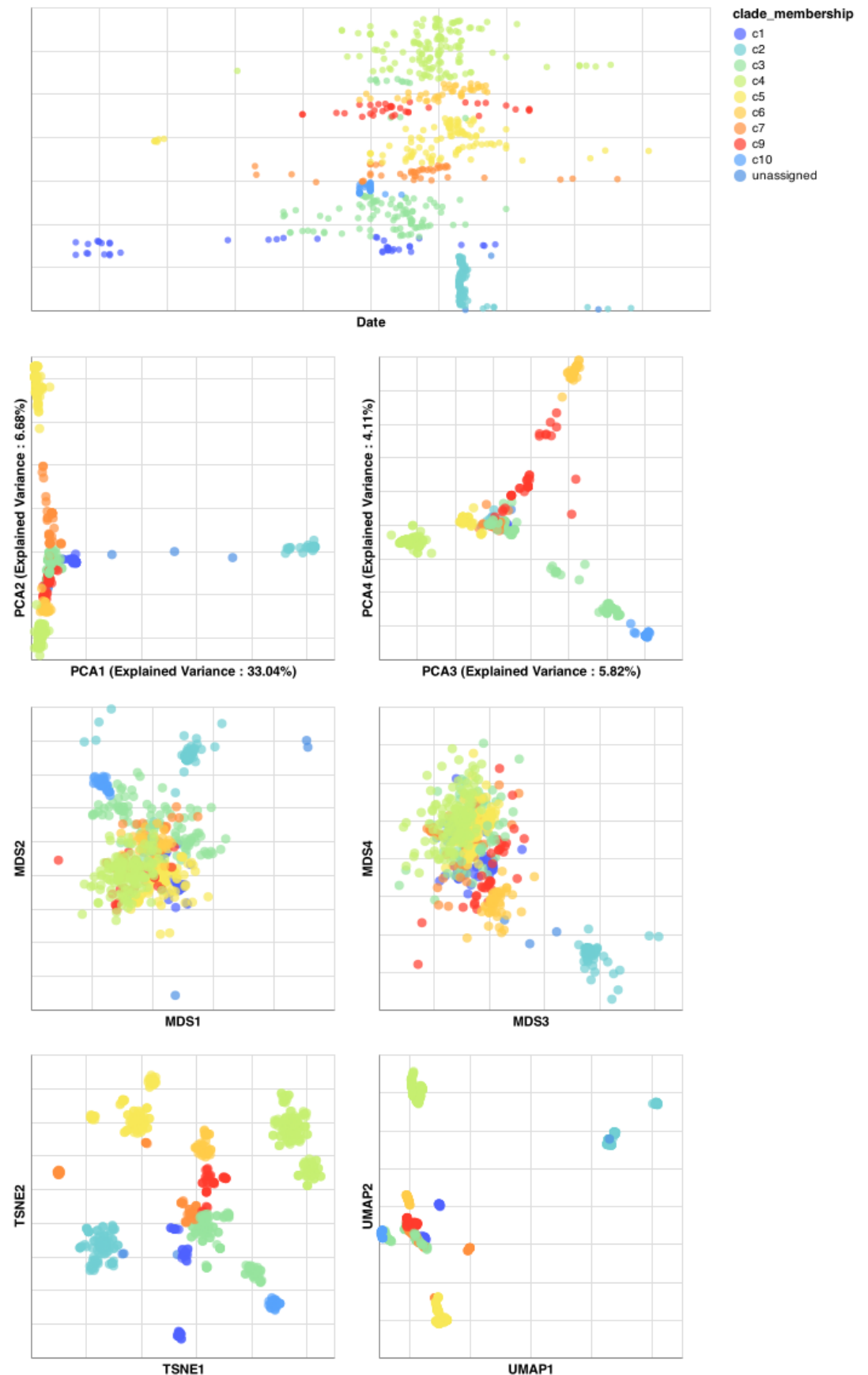
Figure 4. Genetic cartography of Zika strains by dimensionality reduction methods. PCA (components 1 and 2 upper left, components 3 and 4 upper right), MDS (middle left and middle right), t-SNE (lower left), and UMAP (lower right) compared to inferred phylogeny (upper plot). PC1 and PC2 delineate the variance between the Americas and Asia, and PC3 and PC4 are included to better display the variance within the Americas.