

# Genetic cartography reveals ancestral relationships of human pathogenic viruses

Sravani Nanduri<sup>1</sup>, Allison Black<sup>2</sup>, Trevor Bedford<sup>2,3</sup>, John Huddleston<sup>2\*</sup>

**1** Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA

**2** Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

**3** Howard Hughes Medical Institute, Seattle, WA, USA

\* jhuddles@fredhutch.org

## Abstract

[285 words, limit is 300] Public health studies commonly infer phylogenies from viral genome sequences to understand transmission dynamics and identify clusters of genetically-related samples. However, viruses that reassort or recombine violate phylogenetic assumptions and require more sophisticated methods. Even when phylogenies are appropriate, they can be unnecessary. For example, pairwise distances between sequences can be enough to identify clusters of related samples or assign new samples to existing phylogenetic clusters. In this work, we tested whether dimensionality reduction methods could capture known genetic groups within two human pathogenic viruses that cause substantial human morbidity and mortality and frequently reassort or recombine, respectively: seasonal influenza A/H3N2 and SARS-CoV-2. We applied principal component analysis (PCA), multidimensional scaling (MDS), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) to sequences with well-defined phylogenetic clades and either reassortment (H3N2) or recombination (SARS-CoV-2). For each low-dimensional embedding of sequences, we calculated the correlation between

pairwise genetic and Euclidean distances in the embedding and applied a hierarchical clustering method to identify clusters in the embedding. We measured the accuracy of these clusters compared to previously defined phylogenetic clades, reassortment clusters, or recombinant lineages. We found that MDS maintained the strongest correlation between pairwise genetic and Euclidean distances between sequences, best captured the intermediate placement of recombinant lineages between parental lineages, and most accurately identified reassortment groups. Clusters from t-SNE most accurately recapitulated known phylogenetic clades. We show that simple statistical methods without a biological model can accurately represent known genetic relationships for relevant human pathogenic viruses. Our open source implementation of these methods for analysis of viral genome sequences can be easily applied when phylogenetic methods are either unnecessary or inappropriate.

## Author summary

TBD.

## Introduction

Tracking the evolution of human pathogenic viruses in real time enables epidemiologists to respond quickly to emerging epidemics and local outbreaks [1]. Real-time analyses of viral evolution typically rely on phylogenetic methods that can reconstruct the evolutionary history of viral populations from their genome sequences and estimate states of inferred ancestral viruses from the resulting trees including their most likely genome sequence, time of circulation, and geographic location [2–4]. 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 violate this assumption through processes of reassortment or recombination, as seen in seasonal influenza [5, 6] and seasonal coronaviruses [7], respectively. Researchers account for these evolutionary mechanisms by limiting their analyses to individual genes [8, 9], combining multiple genes despite their different evolutionary histories [10], or developing more sophisticated models to represent the

joint likelihoods of multiple co-evolving lineages with ancestral reassortment or  
15 recombination graphs [11, 12]. However, several key questions in genomic epidemiology  
16 do not require full phylogenetic inference of ancestral relationships and states. For  
17 example, genomic epidemiologists commonly need to 1) identify clusters of  
18 closely-related genomes that represent regional outbreaks or new variants of  
19 concern [13–16], 2) place newly sequenced viral genomes in the evolutionary context of  
20 other circulating samples [17–19], and 3) visualize the genetic relationships among  
21 closely related virus samples [20,21]. Given that these common use cases rely on genetic  
22 distances between samples, tree-free statistical methods that operate on pairwise  
23 distances could be sufficient to address each case. As these tree-free methods lack a  
24 formal biological model of evolutionary relationships, they make weak assumptions  
25 about the input data and therefore should be applicable to pathogen genomes that  
26 violate phylogenetic assumptions.  
27

Common statistical approaches to analyzing variation from genome alignments start  
28 by transforming alignments into a matrix that codes each distinct nucleotide character  
29 as an integer or a distance matrix representing the pairwise distances between sequences.  
30 The first of these transformations is the first step prior to performing a principal  
31 component analysis (PCA) to find orthogonal representations of the inputs that explain  
32 the most variance [22]. The second transformation calculates the number of mismatches  
33 between each pair of aligned genome sequences, also known as the Hamming distance,  
34 to create a distance matrix. Most phylogenetic methods begin by building a distance  
35 matrix for all sequences in a given multiple sequence alignment. Dimensionality  
36 reduction algorithms such as multidimensional scaling (MDS) [23], t-distributed  
37 stochastic neighbor embedding (t-SNE) [24], and uniform manifold approximation and  
38 projection (UMAP) [25] accept such distance matrices as an input and produce a  
39 corresponding low-dimensional representation or “embedding” of those data. Both types  
40 of transformation allow us to reduce high-dimensional genome alignments ( $M \times N$   
41 values for  $M$  genomes of length  $N$ ) to low-dimensional embeddings where clustering  
42 algorithms and visualization are more tractable. Additionally, distance-based methods  
43 can reflect the presence or absence of insertions and deletions in an alignment that  
44 phylogenetic methods ignore.  
45

Each of the embedding methods mentioned above has been applied previously to  
46

genomic data to identify clusters of related genomes and visualize relationships between  
47 individuals. Although PCA is a generic linear algebra algorithm that optimizes for an  
48 orthogonal embedding of the data, the principal components from single nucleotide  
49 polymorphisms (SNPs) represent mean coalescent times and therefore recapitulate  
50 broad phylogenetic relationships [26]. PCA has been applied to SNPs of human  
51 genomes [26–29] and to multiple sequence alignments of viral genomes [30]. MDS  
52 attempts to embed input data into a lower-dimensional representation such that each  
53 pair of data points are as far apart in the embedding as they are in the original data.  
54 MDS has been applied to multiple gene segments of seasonal influenza viruses to  
55 visualize evolutionary relationships between segments [31]. Both t-SNE and UMAP  
56 build on manifold learning methods like MDS to find low-dimensional embeddings of  
57 data that place similar points close together and dissimilar points far apart [32]. These  
58 methods have been applied to SNPs from human genomes [33] and single-cell  
59 transcriptomes [34, 35].  
60

Although these methods are commonly used for qualitative studies of evolutionary  
61 relationships, few studies have attempted to quantify patterns observed in the resulting  
62 embeddings and no studies have investigated the value of applying these methods to  
63 human pathogenic viruses. To this end, we tuned and validated the performance of  
64 PCA, MDS, t-SNE, and UMAP with genomes from simulated influenza-like and  
65 coronavirus-like populations and then applied these methods to natural populations of  
66 seasonal influenza virus A/H3N2 and SARS-CoV-2. These natural viruses are highly  
67 relevant as major causes of global human mortality, common subjects of real-time  
68 genomic epidemiology, and representatives of reassortant and recombinant human  
69 pathogens. For each combination of virus and embedding method, we quantified the  
70 relationship between pairwise genetic and Euclidean embedding distances, identified  
71 clusters of closely-related genomes in embedding space, and evaluated the accuracy of  
72 clusters compared to genetic groups defined by experts and biologically-informed  
73 models. Finally, we tested the ability of these methods to identify reassortment of  
74 seasonal influenza virus hemagglutinin (HA) and neuraminidase (NA) segments and  
75 recombination in SARS-CoV-2 genomes. These results inform our recommendations for  
76 future applications of these methods including which are most effective for specific  
77 problems in genomic epidemiology and which parameters researchers should use for each  
78

method.

79

## Materials and methods

80

[This placement of methods before results breaks with PLoS's default organization.  
81  
This organization follows that used by the TreeKnit paper which seemed to be a useful  
82  
model for this paper.]  
83

84

### Embedding methods

84

We selected four standard and common dimensionality reduction (or “embedding”)  
85  
methods to apply to human pathogenic viruses: PCA, MDS, t-SNE, and UMAP. PCA  
86  
operates on a matrix with samples in rows, “features” in columns, and numeric values  
87  
in each cell [22]. To apply PCA to multiple sequence alignments, we transformed each  
88  
nucleotide value into a corresponding integer (A to 1, G to 2, C to 3, T to 4, and all  
89  
other values to 5) and applied scikit-learn’s PCA implementation to the resulting  
90  
numerical matrix with the “full” singular value decomposition solver and 10  
91  
components [36].  
92

93

The remaining three methods operate on a distance matrix. We constructed a  
93  
distance matrix from a multiple sequence alignment by calculating the pairwise  
94  
Hamming distance between nucleotide sequences. By default, the Hamming distance  
95  
only counted mismatches between pairs of standard nucleotide values (A, C, G, and T),  
96  
ignoring other values including gaps. We implemented an optional mode that  
97  
additionally counted each occurrence of consecutive gap characters in either input  
98  
sequence as individual insertion/deletion (“indel”) events.  
99

100

We applied scikit-learn’s MDS implementation to a given distance matrix, with an  
100  
option to set the number of components in the resulting embedding [36]. Similarly, we  
101  
applied scikit-learn’s t-SNE implementation, with options to set the “perplexity” and  
102  
the “learning rate”. The perplexity controls the number of neighbors the algorithm uses  
103  
per input sample to determine an optimal embedding [24]. This parameter effectively  
104  
determines the balance between maintaining “local” or “global” structure in the  
105  
embedding [35]. The learning rate controls how rapidly the t-SNE algorithm converges  
106  
on a specific embedding [24,37] and should scale with the number of input samples [38].  
107

We initialized t-SNE embeddings with the first two components of the corresponding  
108 PCA embedding, as previously recommended to obtain more accurate global  
109 structure [32, 35]. Finally, we applied the *umap-learn* Python package written by  
110 UMAP’s authors, with options to set the number of “nearest neighbors” and the  
111 “minimum distance” [25]. As with t-SNE’s perplexity parameter, the nearest neighbors  
112 parameter determines how many adjacent samples the UMAP algorithm considers per  
113 sample to find an optimal embedding. The minimum distance sets the lower limit for  
114 how close any two samples can map next to each other in a UMAP embedding. Lower  
115 minimum distances allow tighter groups of samples to form. For both t-SNE and  
116 UMAP, we used the default number of components of 2.  
117

### Simulation of influenza-like and coronavirus-like populations

118

Given the relative lack of prior application of dimensionality reduction methods to  
119 human pathogenic viruses, we first attempted to understand the behavior and optimal  
120 parameter values for these methods when applied to simulated viral populations with  
121 well-defined evolutionary parameters. To this end, we simulated populations of  
122 influenza-like and coronavirus-like viruses using SANTA-SIM [39]. These simulated  
123 populations allowed us to identify optimal parameters for each embedding method,  
124 without overfitting to the limited data available for natural viral populations. For each  
125 population type described below, we simulated five independent replicates with fixed  
126 random seeds for over 55 years, filtered out the first 10 years of each population as a  
127 burn-in period, and analyzed the remaining years.  
128

We simulated influenza-like populations as previously described with 1,700 bp  
129 hemagglutinin sequences [40]. As in that previous study, we scaled the number of  
130 simulated generations per real year to 200 per year to match the observed mutation rate  
131 for natural H3N2 HA sequences, and we sampled 10 genomes every 4 generations for  
132 12,000 generations (or 60 years of real time).  
133

We simulated coronavirus-like populations as previously described for human  
134 seasonal coronaviruses with genomes of 21,285 bp [12]. For the current study, we  
135 assigned 30 generations per real year to obtain mutation rates similar to the  $8 \times 10^{-4}$   
136 substitutions per site per year estimated for SARS-CoV-2 [41]. To account for the effect  
137

of recombination on optimal method parameters, we simulated populations with a recombination rate of  $10^{-5}$  events per site per year based on human seasonal coronaviruses for which recombination rates are well-studied [12, 42]. We calibrated the overall recombination probability in SANTA-SIM such that the number of observed recombination events per year matched the expected number for human seasonal coronaviruses (0.3 per year) [12]. To assist with this calibration of recombination events per year, we modified the SANTA-SIM source code to emit a boolean status of “is recombinant” for each sampled genome. This change allowed us to identify recombinant genomes by their metadata in downstream analyses and calculate the number of recombination events observed per year. For each replicate population, we sampled 15 genomes every generation for 1,700 generations (or approximately 56 years of real time).

## Optimization of embedding method parameters

We identified optimal parameter values for each embedding method with time series cross-validation of embeddings based on simulated populations [43]. To increase the interpretability of embedding space, we defined parameters as “optimal” when they maximized the linear relationship between pairwise genetic distance of viral genomes and the corresponding Euclidean distance between those same genomes in an embedding. This optimization approach allowed us to also determine the degree to which each method could recapitulate this linear relationship.

For each simulated population replicate, we created 10 training and test datasets that each consisted of 4 years of training data and 4 years of test data preceded by a 1-year gap from the end of the training time period. These settings produced training/test data with 2000 samples each for influenza-like populations and 1800 samples each for coronavirus-like populations. For each combination of training/test dataset, embedding method, and method parameters, we applied the following steps. We created an embedding from the training data with the given parameters, fit a linear model to estimate pairwise genetic distance from pairwise Euclidean distance in the embedding, created an embedding from the test data, estimated the pairwise genetic distance for genomes in the test data based on their Euclidean distances and the linear model fit to the training data, and calculated the mean absolute error (MAE) between

estimated and observed genetic distances in the test data. We summarized the error for  
168 a given population type, method, and method parameters across all population  
replicates and training/test data by calculating the median of the MAE. For all method  
parameters except those controlling the number of components used for the embedding,  
169 we selected the optimal parameters as those that minimized the median MAE for a  
given embedding method. Since increasing the number of components used by PCA and  
170 MDS allows these methods to overfit to available data, we selected the optimal number  
of components for these methods as the number beyond which the median MAE did not  
171 decrease by at least 1 nucleotide. This approach follows the same concept from the  
MDS algorithm itself where optimization occurs iteratively until the algorithm reaches a  
172 predefined error threshold [23].  
173

With the approach described above, we tested each method across a range of  
relevant parameters with all combinations of parameter values. For PCA, we tested the  
number of components between 2 and 6. For MDS, we tested the number of  
components between 2 and 10. [The difference in number of components between PCA  
and MDS sticks out here. We should use the same number for both or justify using  
different numbers.] For t-SNE, we tested perplexity values of 15, 30, 100, 200, and 300,  
and we tested learning rates of 100, 200, and 500. For UMAP, we tested nearest  
neighbor values of 25, 50, and 100, and we tested values for the minimum distance that  
points can be in an embedding of 0.05, 0.1, and 0.25.  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187

## Selection of natural virus population data

We selected recent publicly available genome sequences and metadata for seasonal  
influenza H3N2 HA and NA genes and SARS-CoV-2 genomes from INSDC  
databases [44]. For both viruses, we divided the available data into “early” and “late”  
datasets to use as training and test data, respectively, for identification of virus-specific  
clustering parameters. [First mention of clustering happens here before we define what  
clustering is later on. Maybe ok as long as we reference the “later on” bit here  
parenthetically?]  
188  
189  
190  
191  
192  
193  
194  
195

For analyses that focused only on H3N2 HA data, we defined the early dataset  
between January 2016 and January 2018 and the late dataset between January 2018 to  
196  
197

January 2020. These datasets reflected two years of recent H3N2 evolution up to the  
198 time when the SARS-CoV-2 pandemic disrupted seasonal influenza circulation. For both  
199 early and late datasets, we evenly sampled 25 sequences per country, year, and month,  
200 excluding known outliers. With this sampling scheme, we selected 1,523 HA sequences  
201 for the early dataset and 1,073 for the late dataset. For analyses that combined H3N2  
202 HA and NA data, we defined a single dataset between January 2016 and January 2018,  
203 keeping 1,607 samples for which both HA and NA have been sequenced.  
204

For SARS-CoV-2 data, we defined the early dataset between January 1, 2020 and  
205 January 1, 2022 and the late dataset between January 1, 2022 and November 3, 2023.  
206 For the early dataset, we evenly sampled 1,736 SARS-CoV-2 genomes by geographic  
207 region, year, and month, excluding known outliers. For the late dataset, we used the  
208 same even sampling by space and time to select 1,309 representative genomes. In  
209 addition to these genomes, we sampled at most 20 genomes per Nextclade pango lineage  
210 for 10 known recombinant lineages (XAY, XBB, XBB.1, XBC, XBF, XBL, XC, XD,  
211 XE, XF, and XG) and their corresponding parental lineages (AY.29, AY.4, AY.45,  
212 B.1.1.7, B.1.617, BA.1, BA.2, BA.2.75, BA.4, BA.5, BA.5.2.3, BJ.1, BM.1.1.1, and  
213 CJ.1) as defined by <https://libguides.mskcc.org/SARS2/recombination>. [At this point,  
214 we haven't defined "Pango lineages" yet, but I don't know that it makes sense to define  
215 lineages in this section. Curious what other people think.] With these additional  
216 genomes, the late SARS-CoV-2 dataset included 1,668 total genomes.  
217

## Evaluation of linear relationships between genetic distance and Euclidean distance in embeddings

To evaluate the biological interpretability of distances between samples in  
220 low-dimensional embeddings, we plotted the pairwise Euclidean distance between  
221 samples in each embedding against the corresponding genetic distance between the same  
222 samples. We calculated Euclidean distance using all components of the given embedding  
223 (e.g., 2 components for PCA, t-SNE, and UMAP and 3 components for MDS). For each  
224 embedding, we fit a linear model between Euclidean and genetic distance and calculated  
225 the squared Pearson's correlation coefficient,  $R^2$ . The distance plots provide a  
226 qualitative assessment of each embedding's local and global structure relative to a  
227

biologically meaningful scale of genetic distance, while the linear models and correlation  
coefficients quantify the global structure in the embeddings.

## Phylogenetic analysis

For each natural population described above, we created an annotated, time-scaled phylogenetic tree. For seasonal influenza H3N2 HA and NA sequences, we aligned sequences with MAAFT (version 7.486) [45, 46] using the *augur align* command (version 22.0.3) [47]. For SARS-CoV-2 sequences, we used existing reference-based alignments provided by the Nextstrain team ([https://docs.nextstrain.org/projects/ncov/en/latest/reference/remote\\_inputs.html](https://docs.nextstrain.org/projects/ncov/en/latest/reference/remote_inputs.html)) and generated with Nextalign (version 2.14.0) [19]. We inferred a phylogeny with IQ-TREE (version 2.1.4-beta) [48] using the *augur tree* command and inferred a time tree with TreeTime (version 0.10.1) [4] using the *augur refine* command. We visualized phylogenies with Auspice [49], after first converting the trees to Auspice JSON format with *augur export*.

## Definitions of genetic groups by experts or biologically-informed models

We annotated phylogenetic trees with genetic groups previously identified by experts or assigned by biologically-informed models. For seasonal influenza H3N2, the World Health Organization assigns “clade” labels to clades in HA phylogenies that appear to be genetically or phenotypically distinct from other recently circulating H3N2 samples. We used the latest clade definitions for H3N2 maintained by the Nextstrain team as part of their seasonal influenza surveillance efforts [50].

As seasonal influenza clades only account for the HA gene and lack information about reassortment events, we assigned joint HA and NA genetic groups using a biologically-informed model, TreeKnit [11]. TreeKnit infers ancestral reassortment graphs from two gene trees, finding groups of samples for which both genes share the same history. These groups, also known as maximally compatible clades (MCCs), represent samples whose HA and NA genes have reassorted together. TreeKnit attempts to resolve polytomies in one tree using information present in the other tree(s). Input

trees for TreeKnit must contain the same samples and root on the same sample. Because 257  
of these TreeKnit expectations, we inferred HA and NA trees with IQ-TREE with a 258  
custom argument to collapse near-zero-length branches ('-czb'). We rooted the resulting 259  
trees on the same sample that we used as an alignment reference, A/Beijing/32/1992, 260  
and pruned this sample prior to downstream analyses. We applied TreeKnit to the 261  
rooted HA and NA trees with a gamma value of 2.0 and the ‘–better-MCCs’ flag, as 262  
previously recommended for H3N2 analyses [11]. Finally, we filtered the MCCs 263  
identified by TreeKnit to retain only those with at least 10 samples and to omit the root 264  
MCC that represented the most recent common ancestor in both HA and NA trees. 265

For SARS-CoV-2, we used both expert-defined “Nextstrain clades” [51–53] and 266  
computationally-defined Pangolin lineages [17] provided by Nextclade as “Nextclade 267  
pango” annotations. Nextstrain clade definitions represent the World Health 268  
Organization’s variants of concern and other phylogenetic clades that have reached 269  
minimum global and regional frequencies and growth rates. Pangolin lineages represent 270  
a combination of lineages assigned by a machine learning model (pangoLEARN) and 271  
expert-curated lineages (<https://github.com/cov-lineages/pango-designation>) and must 272  
contain at least 5 samples with an unambiguous evolutionary event. As such, Nextstrain 273  
clades represent a much coarser genetic resolution than Pangolin lineages. Additionally, 274  
Pangolin lineages produced by recombination receive a lineage name prefixed by an “X”, 275  
while Nextstrain clades do not explicitly reflect recombination events. 276

Since Pangolin lineages can represent much smaller genetic groups than are 277  
practically useful, we collapsed lineages with fewer than 10 samples in our analysis into 278  
their parental lineages using the pango\_aliasor tool 279  
([https://github.com/corneliusroemer/pango\\_aliasor](https://github.com/corneliusroemer/pango_aliasor)). Specifically, we counted the 280  
number of samples per lineage, sorted lineages in ascending order by count, and 281  
collapsed each lineage with a count less than 10 into its parental lineage in the 282  
count-sorted order. This approach allowed small lineages to aggregate with other small 283  
parental lineages and meet the 10-sample threshold. We used these “collapsed 284  
Nextclade pango” lineages for subsequent analyses. 285

## Clustering of samples in embeddings

286

To understand how well embeddings of genetic data could capture previously defined genetic groups, we applied an unsupervised clustering algorithm, HDBSCAN [54], to each embedding. HDBSCAN identifies initial clusters from high-density regions in the input space and merges these clusters hierarchically. This algorithm allowed us to avoid defining an arbitrary or biased expected number of clusters *a priori*. HDBSCAN provides parameters to tune the minimum number of samples required to seed an initial cluster (“min samples”), the minimum size for a final cluster (“min size”), and the minimum distance between initial clusters below which those clusters are hierarchically merged (“distance threshold”). We hardcoded the min samples to 5 to minimize the number of spurious initial clusters and min size to 10 to reflect our interest in genetic groups with at least 10 samples throughout our analyses. HDBSCAN calculates the distance between clusters on the Euclidean scale of each embedding. To account for embedding-specific distances, we performed a coarse grid search of distance threshold values for each virus type and embedding method.

300

We performed the grid search on the early datasets for both seasonal influenza H3N2 HA and SARS-CoV-2. For each dataset and embedding method, we applied HDBSCAN clustering with a distance threshold between 0 and 7 inclusive with steps of 0.5 between values. For a given threshold, we obtained sets of samples assigned to HDBSCAN clusters from the embedding. We evaluated the accuracy of these clusters with variation of information (VI) which calculates the distance between two sets of clusters of the same samples [55]. When two sets of clusters are identical, VI equals 0. When the sets are maximally different, VI is  $\log N$  where  $N$  is the total number of samples. To make VI values comparable across datasets, we normalized each value by dividing by  $\log N$ , following the pattern used to validate TreeKnit’s MCCs [11]. Unlike other standard metrics like accuracy, sensitivity, or specificity, VI distances do not favor methods that tend to produce more, smaller clusters. For each virus dataset and embedding method, we identified the distance threshold that minimized the normalized VI between HDBSCAN clusters and genetic groups defined by experts or biologically-informed models (“Nextstrain clade” for seasonal influenza and both “Nextstrain clade” and “collapsed Nextclade pango lineage” for SARS-CoV-2). HDBSCAN allows samples to not

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

belong to a cluster and assigns these samples a numeric label of -1. We intentionally  
317 included all unassigned samples in the normalized VI calculation thereby penalizing  
318 cluster parameters that increased the number of unassigned samples by increasing their  
319 VI values. Finally, we used these optimal distance thresholds to identify clusters in  
320 out-of-sample data from the late datasets for both viruses and calculate the normalized  
321 VI between those clusters and previously defined genetic groups.  
322

## Evaluating robustness of embedding cluster accuracy

323

The cluster accuracies we estimated for late H3N2 HA and SARS-CoV-2 datasets  
324 represented a single VI measurement for a single pathogen dataset. To understand how  
325 robust these accuracies were across different datasets, we generated alternate random  
326 samples from both late pathogen datasets using the same geographic and temporal  
327 grouping but with different numbers of sequences per group and different random seeds.  
328 Specifically, we sampled 5, 10, 15, 20, or 25 sequences per group for five replicates per  
329 pathogen (random seeds of 0, 1, 2, 3, and 4), embedded these sequences with each  
330 method, identified clusters in embeddings, and calculated the VI distance between those  
331 clusters and Nextstrain clade assignments. We plotted the distribution of the resulting  
332 VI distances, to estimate the variance of these values caused by sampling bias.  
333

## Identification of cluster-specific mutations

334

To better understand the genetic basis of embedding clusters, we identified  
335 cluster-specific mutations for all HDBSCAN clusters. First, we found all mutations  
336 between each sample's sequence and the reference sequence used to produce the  
337 alignment, considering only A, C, G, T, and gap characters. Within each cluster, we  
338 identified mutations that occurred in at least 10 samples and in at least 50% of samples  
339 in the cluster. We recorded the resulting mutations per cluster in a table with columns  
340 for the embedding method, the position of the mutation, the derived allele of the  
341 mutation, and a list of the distinct clusters the mutation appeared in. From this table,  
342 we could identify mutations that only occurred in specific clusters and mutations that  
343 distinguished sets of clusters from each other.  
344

## Assessment of HA/NA reassortment in seasonal influenza

345

### populations

346

To assess the ability of embedding methods to detect reassortment in seasonal influenza populations, we applied each method to either HA alignments only or concatenated alignments of HA and NA sequences from the same samples, performed HDBSCAN clustering with the optimal distance threshold for the given method, and calculated the normalized VI between the resulting clusters and TreeKnit MCCs. To minimize the effects of missing data on the PCA embeddings, we dropped all columns with N characters from the HA and HA/NA alignments prior to producing PCA embeddings. We used the original alignments to calculate distance matrices for all other methods, since distance-based methods can ignore N characters in pairwise comparisons. We compared normalized VI values for the HA-only clusters of each method to the corresponding VI values for the HA/NA clusters. Lower VI values in the HA/NA clusters than HA-only clusters indicated better clustering of samples into known reassortment groups.

359

## Assessment of recombination in SARS-CoV-2 populations

360

To assess the ability of embedding methods to detect recombination in late SARS-CoV-2 populations (2022-2023), we calculated the Euclidean distances in low-dimensional space between the 10 known recombinant lineages and their respective parental lineages described in “Selection of natural virus population data” above. Given that we optimized each method’s parameters to maximize a linear relationship between genetic and Euclidean distance, we expected embeddings to place recombinant lineages between their parental lineages, reflecting the intermediate genetic state of the recombinants. For a recombinant lineage  $X$  and its parental lineages  $A$  and  $B$ , we calculated the average pairwise Euclidean distance,  $D$ , between samples in  $A$  and  $B$ ,  $A$  and  $X$ , and  $B$  and  $X$ . We identified lineages that mapped properly as those for which  $D(A, X) < D(A, B)$  and  $D(B, X) < D(A, B)$ . We also identified lineages for which the recombinant lineage placed closer to at least one parent than the distance between the parents. Note that we used the original uncollapsed “Nextclade pango” annotations to identify samples in each lineage, as these were the lineage names used to include recombinant samples in the

374

analysis and define known relationships between recombinant and parental lineages.

375

## Data and software availability

376

The entire workflow for our analyses was implemented with Snakemake [56]. We have  
377 provided all source code, configuration files, and datasets at  
378

<https://github.com/blab/cartography>. Interactive phylogenetic trees and corresponding  
379 embeddings for natural populations are available at  
380

<https://nextstrain.org/community/blab/cartography/>. The *pathogen-embed* Python  
381 package, available at <https://pypi.org/project/pathogen-embed/>, provides command  
382 line utilities to calculate distance matrices (*pathogen-distance*), calculate embeddings  
383 per method (*pathogen-embed*), and apply hierarchical clustering to embeddings  
384 (*pathogen-cluster*).  
385

## Results

386

### Simulated populations enable tuning of embedding method

387

#### parameters

388

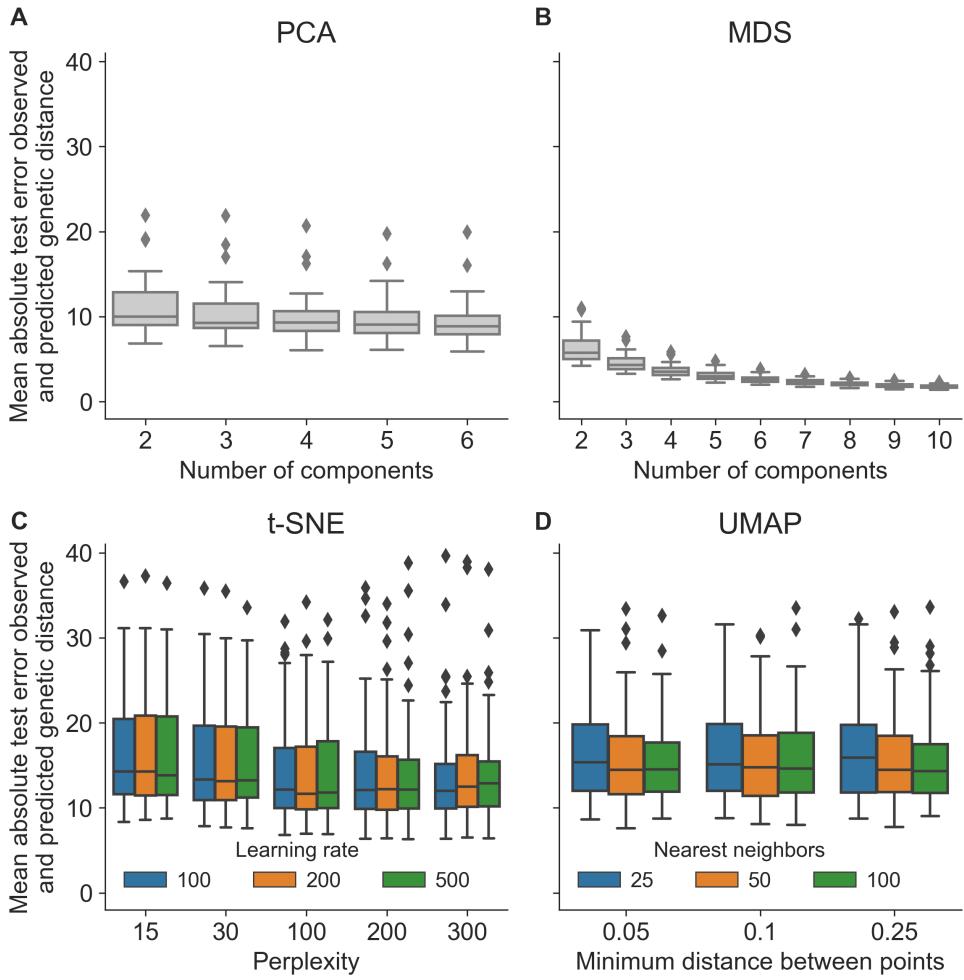
To understand how well PCA, MDS, t-SNE, and UMAP could represent genetic  
389 relationships between samples of human pathogen viruses under well-defined  
390 evolutionary conditions, we simulated influenza-like and coronavirus-like populations,  
391 created embeddings for each population across a range of method parameters, and  
392 identified optimal parameters as those that maximized a linear relationship between  
393 genetic distance and Euclidean distance in low-dimensional space (see Methods).  
394 Specifically, we selected parameters that minimized the median of the mean absolute  
395 error (MAE) between observed pairwise genetic distances of simulated genomes and  
396 predicted genetic distances for those genomes based on their Euclidean distances in each  
397 embedding. For methods like PCA and MDS where increasing the number of  
398 components available to the embedding could lead to overfitting, we selected the  
399 maximum number of components beyond which the median MAE did not decrease by  
400 more than 1 nucleotide.  
401

For influenza-like populations, the optimal parameters were 2 components for PCA,  
402

3 components for MDS, perplexity of 100 and learning rate of 200 for t-SNE, and  
403 nearest neighbors of 100 and minimum distance of 0.25 for UMAP. As expected,  
404 increasing the number of components for PCA and MDS gradually decreased the  
405 median MAEs of their embeddings (S1 Fig A and B). However, beyond 2 and 3  
406 components, respectively, the reduction in error did not exceed 1 nucleotide. This result  
407 suggests that there were diminishing returns for the increased complexity of additional  
408 components. Both t-SNE and UMAP embeddings produced a wide range of errors (the  
409 majority between 10 and 20 average mismatches) across all parameter values (S1 Fig C  
410 and D). Embeddings from t-SNE appeared robust to variation in parameters, with a  
411 slight improvement in median MAE associated with perplexity of 100 and little benefit  
412 to any of the learning rate values (S1 Fig C). [Based on these results, we should consider  
413 setting the learning rate to the default for scikit-learn which scales the rate with the  
414 input sample size.] Similarly, UMAP embeddings were robust across the range of tested  
415 parameters, with the greatest benefit coming from setting the nearest neighbors greater  
416 than 25 and no benefit from changing the minimum distance between points (S1 Fig D).  
417

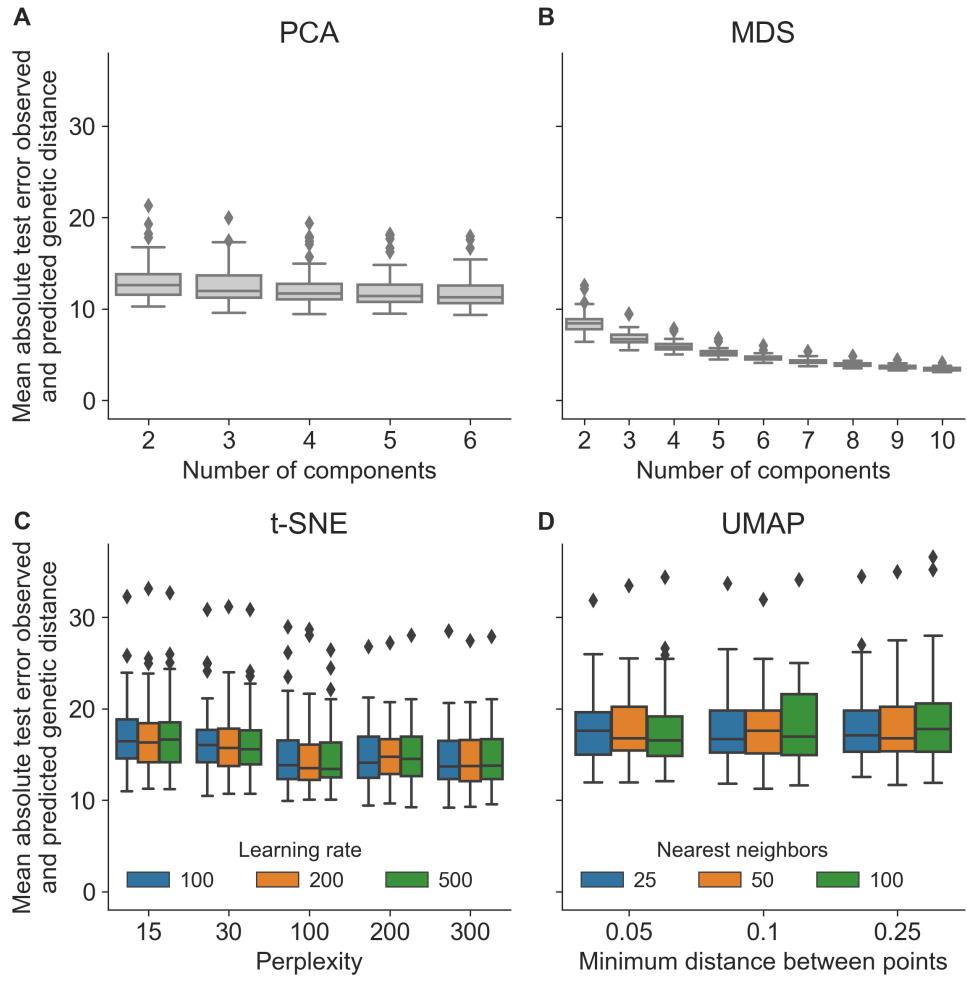
The optimal parameters for coronavirus-like populations were nearly the same as  
418 those for the influenza-like populations. The optimal parameters were 2 components for  
419 PCA, 3 for MDS, perplexity of 100 and learning rate of 500 for t-SNE, and nearest  
420 neighbors of 100 and minimum distance of 0.05 for UMAP. As with influenza-like  
421 populations, both PCA and MDS showed diminishing benefits of increasing the number  
422 of components (S2 Fig A and B). Similarly, we observed little improvement in MAEs  
423 from varying t-SNE and UMAP parameters (S2 Fig C and D). The most noticeable  
424 improvement came from setting t-SNE's perplexity to 100 (S2 Fig C). These results  
425 indicate the limits of t-SNE and UMAP to represent global genetic structure, at least  
426 across the parameter regimes considered here. [An obvious follow-up question would be  
427 whether we can improve MAEs for these methods by increasing components available to  
428 them, too.]  
429

We inspected representative embeddings based on the optimal parameters above for  
430 the first four years of influenza- and coronavirus-like populations. Simulated sequences  
431 collected from the same time period tended to map closer in embedding space,  
432 indicating the maintenance of “local” genetic structure in the embeddings (Fig. 1).  
433 Most embeddings also represented some form of global structure, with later generations  
434



**S1 Fig. Distribution of mean absolute errors (MAE) between observed and predicted pairwise genetic distances per embedding method parameters for simulated influenza-like populations.**

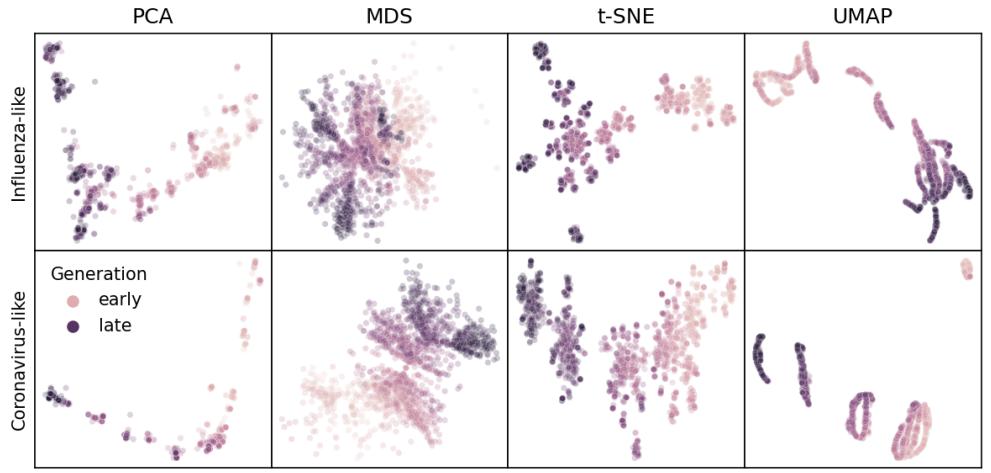
mapping closer to intermediate generations than earlier generations. MDS maintained the greatest continuity between generations for both population types (S3 Fig). In contrast, PCA, t-SNE, and UMAP all demonstrated tighter clusters of samples separated by potentially arbitrary space. The UMAP embedding for the coronavirus-like samples was most extreme in this respect, with a tight cluster of early samples placing far away from all other samples in the embedding including those from nearby generations. These qualitative results matched our expectations based on how well each method maximized a linear relationship between genetic and Euclidean distances during parameter optimization (S1 Fig and S2 Fig).



**S2 Fig. Distribution of mean absolute errors (MAE) between observed and predicted pairwise genetic distances per embedding method parameters for simulated coronavirus-like populations.**

### Embedding clusters recapitulate phylogenetic clades for seasonal influenza H3N2

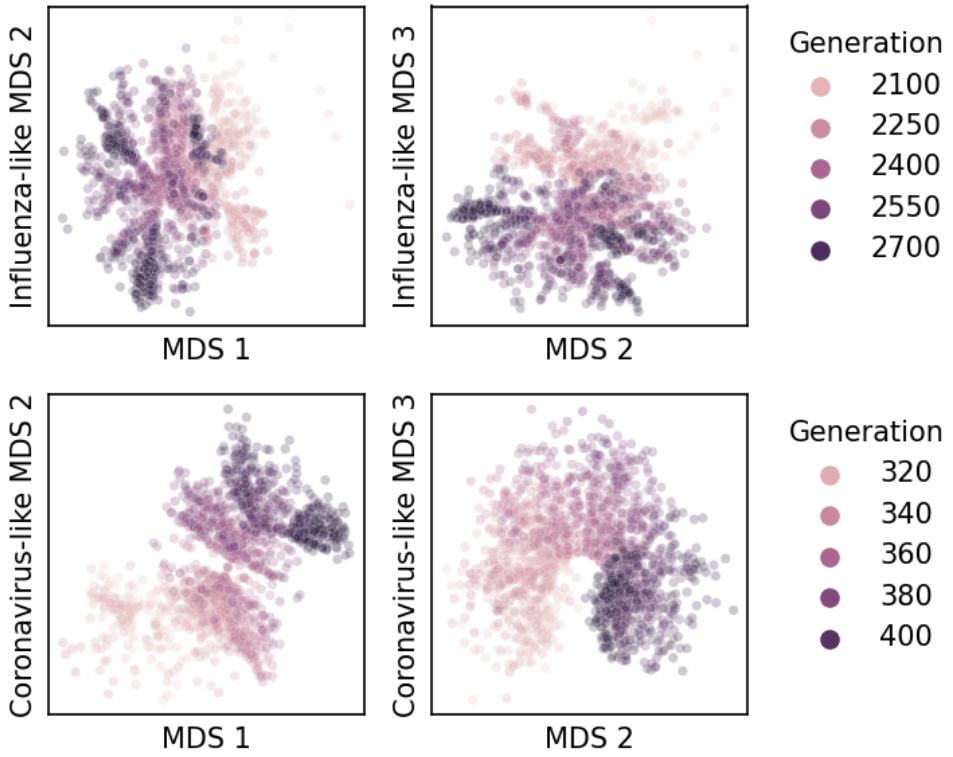
Seasonal influenza H3N2's hemagglutinin (HA) sequences provide an ideal positive control to test embedding methods and clustering in low-dimensional space. H3N2's HA protein evolves rapidly, accumulating amino acid mutations that enable escape from adaptive immunity in human populations [57]. These mutations 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 [58] and submits



**Fig 1. Representative embeddings for simulated populations using optimal parameters per pathogen (rows) and embedding method (columns).** Each panel shows the embedding for sequences from the first four years of a single replicate population for the corresponding pathogen type. Each point represents a simulated viral sequence colored by its generation with darker values representing later generations. The MDS embedding shows the first two of three total components. S3 Fig shows the full MDS embedding for all components.

these sequences to public INSDC databases like NCBI's GenBank [44]. These factors, coupled with HA's relatively short gene size of 1,701 nucleotides, facilitate real-time genomic epidemiology of H3N2 [50] and rapid analysis by the embedding methods we wanted to evaluate.

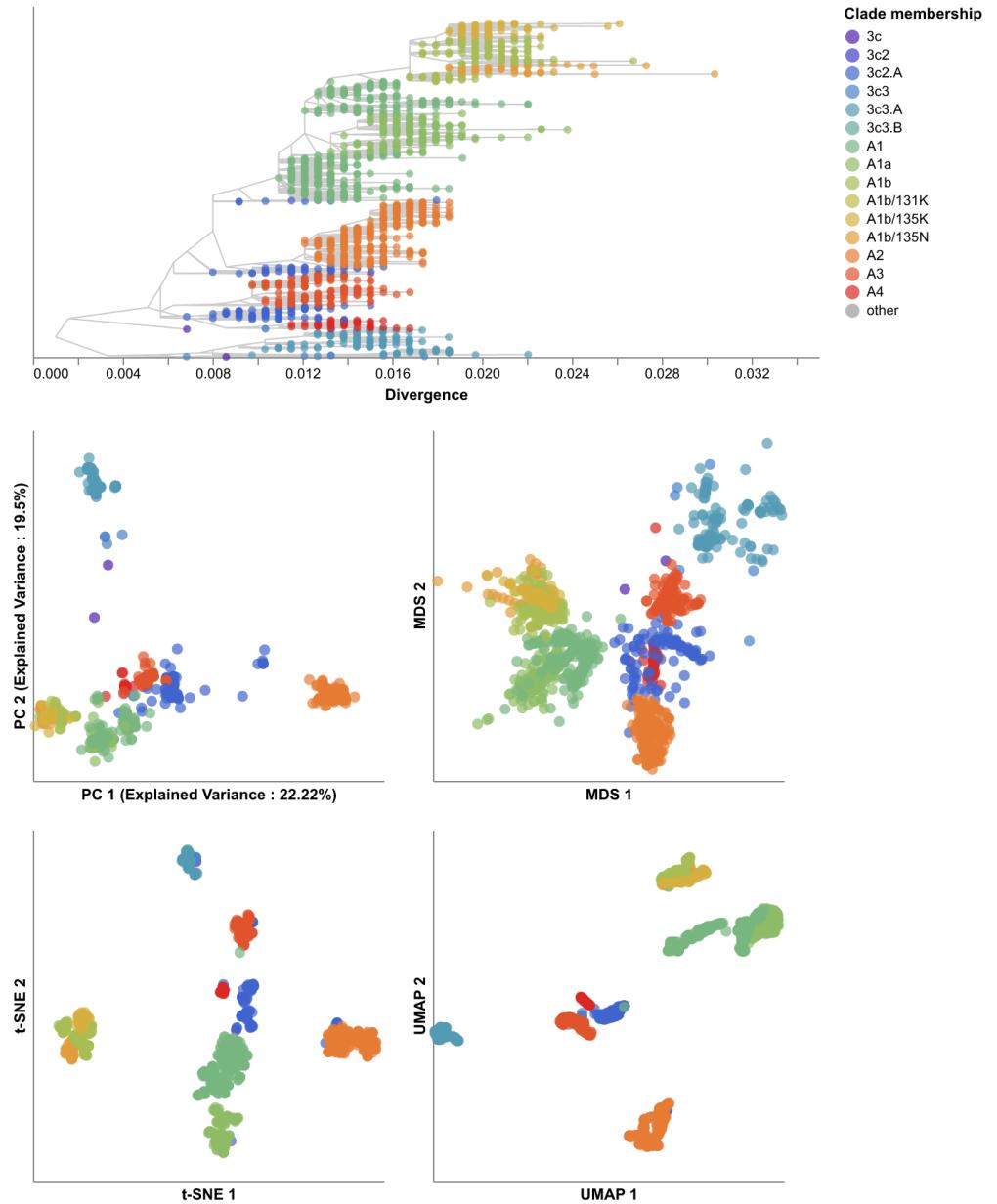
We first applied each embedding method to the early H3N2 HA sequences collected from 2016 through 2018, colored samples by previously defined phylogenetic clades, and inspected the placement of these samples in the embeddings and corresponding phylogeny. All four embedding methods qualitatively recapitulated clade-level groupings observed in the phylogeny (Fig 2). Samples from the same clade generally grouped tightly together. Most embedding methods also clearly delineated larger phylogenetic clades, placing clades A1, A2, A3, A4, and 3c3.A into separate locations in the embeddings. One exception to this pattern was the PCA embedding which grouped samples from clades A3 and A4 into the same space. Despite maintaining local and broader global structure, not all embeddings captured intermediate genetic structure. For example, clade A1b descended from clade A1 and diversified into the smaller subclades A1b/131K, A1b/135K, and A1b/135N. All methods placed A1b far from its ancestor A1, but all methods also placed descendants of A1b into tight clusters together.



**S3 Fig. Representative MDS embeddings for simulated populations using optimal parameters per pathogen (rows) and showing all three components.**

The t-SNE embedding created separate clusters of the three descendant clades, but these clusters all placed so close together in the embedding space that, without previously defined clade labels, we would have visually grouped these samples into a single cluster. These results qualitatively replicate the patterns we observed in embeddings for simulated influenza-like populations (Fig 1).

To quantify the apparent maintenance of local and global structure by all four embedding methods, we calculated the relationship between pairwise genetic and Euclidean distance of samples in each embedding. All four methods maintained a linear relationship between genetic and Euclidean distances for samples that differed by no more than  $\approx 10$  nucleotides (Fig 3). However, only MDS consistently maintained that linearity as genetic distance increased (Pearson's  $R^2 = 0.94$ ). Values of Euclidean distances in MDS corresponded nearly perfectly with values of genetic distances. In contrast, we observed a nonlinear relationship for samples with more genetic differences in PCA (Pearson's  $R^2 = 0.67$ ), t-SNE (Pearson's  $R^2 = 0.37$ ), and UMAP (Pearson's

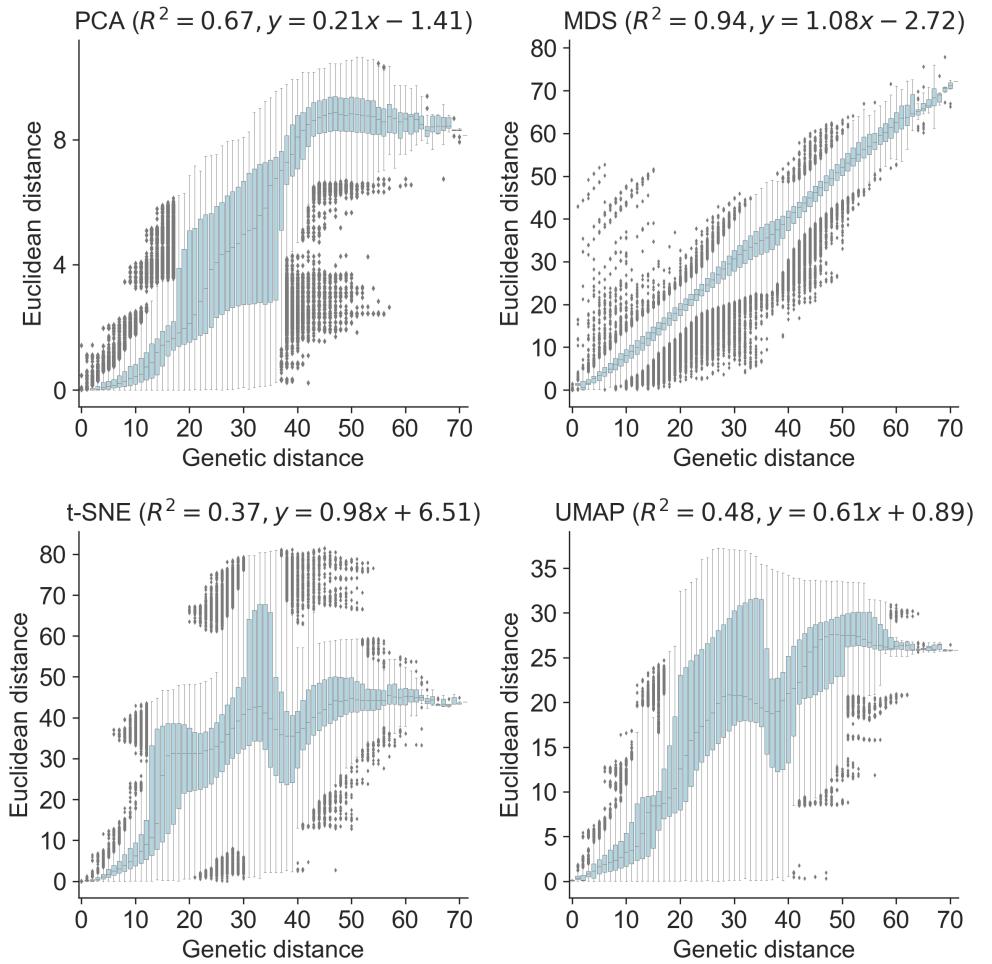


**Fig 2.** Phylogeny of early (2016–2018) influenza H3N2 HA sequences plotted by nucleotide substitutions per site on the x-axis (top) and low-dimensional embeddings of the same sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips in the tree and embeddings are colored by their Nextstrain clade assignment.

$R^2 = 0.48$ ) embeddings. Although the most genetically distant samples mapped far from each other in all of these embeddings, samples with intermediate distances could map much closer or farther than expected by a linear model. In t-SNE and UMAP embeddings, some pairs of samples with intermediate distances of 30–40 nucleotides

mapped farther apart than pairs of samples with much greater genetic distances.

488



**Fig 3. Relationship between pairwise genetic and Euclidean distances in embeddings of early (2016–2018) influenza H3N2 HA sequences by PCA (upper left), MDS (upper right), t-SNE (lower left), and UMAP (lower right). Each boxplot represents the distribution of pairwise Euclidean distances at a given genetic distance.**

Next, we measured how well clusters of H3N2 HA samples in each embedding corresponded to previously defined genetic groups. For each embedding, we assigned cluster labels to each sample with the hierarchical clustering algorithm, HDBSCAN, which does not require an expected number of clusters as input [54]. HDBSCAN does not require definition of a minimum distance that its initial clusters must be from each other to avoid being merged into the same cluster. This distance corresponds to the Euclidean distance between clusters in embedding space which varies by method (Fig 3). To find the optimal minimum distance for HDBSCAN clusters of H3N2 HA data, we

489

490

491

492

493

494

495

496

assigned clusters to each embedding for a range of distance values (0-7) with a step size  
 497 of 0.5 and calculated the accuracy of clusters at each distance value compared to the  
 498 Nextstrain clade assignments shown in Fig 2. We selected the minimum distance value  
 499 per method that minimized the difference between HDBSCAN clusters and clade  
 500 assignments as measured by the normalized variation of information (VI) metric [55]  
 501 (see Methods). The optimal minimum distances were 0.5 for PCA, 3.5 for MDS, 2.0 for  
 502 t-SNE, and 1.0 for UMAP (Table 1). Since Euclidean distances for MDS correspond  
 503 directly to genetic distances, these results show that clusters must be at least 3.5  
 504 nucleotides apart to be considered distinct.  
 505

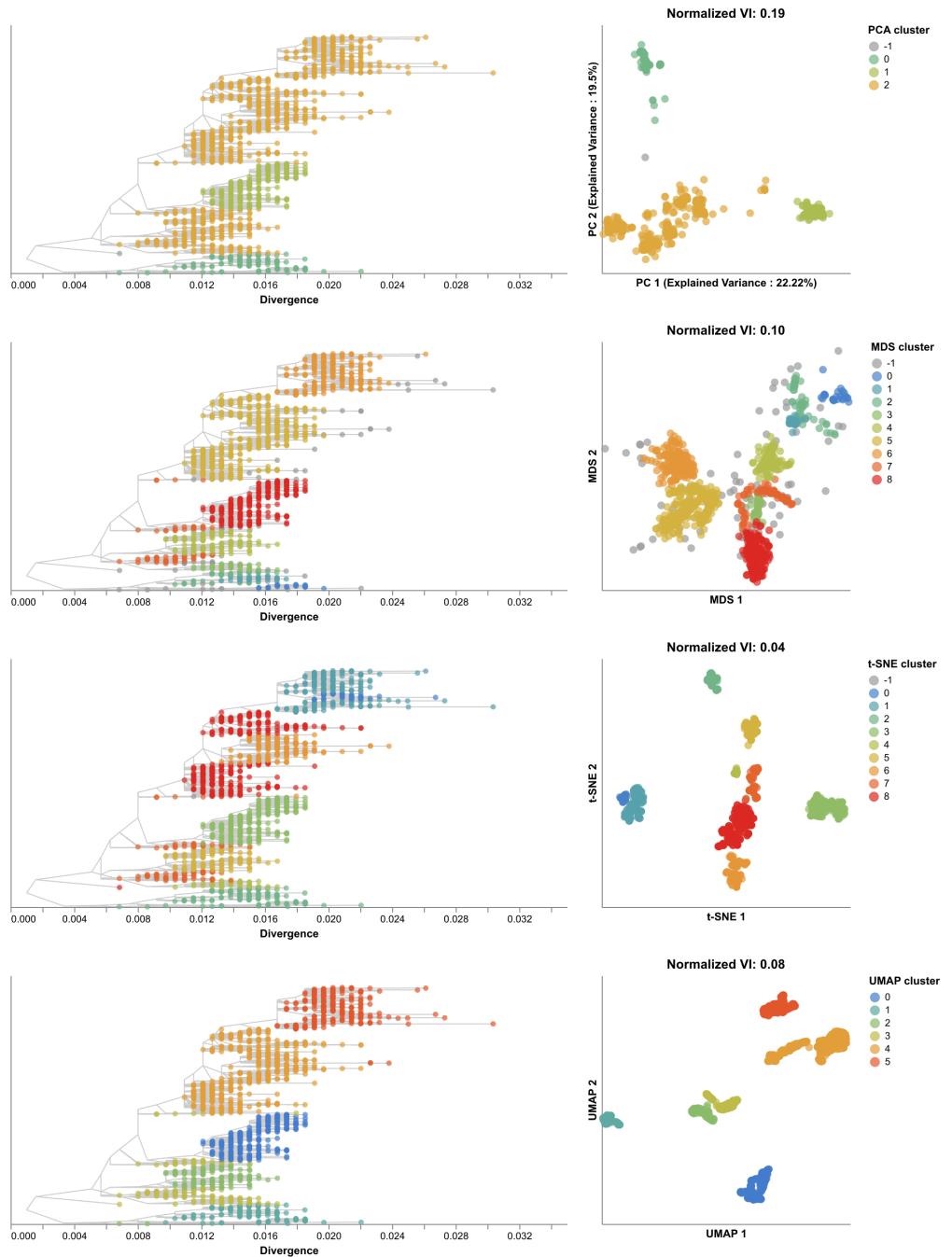
**Table 1. Accuracy of embedding methods per human pathogenic virus sorted by normalized variation of information (VI) distance.** Smaller VI values indicate smaller distances between HDBSCAN clusters and known genetic groups with 0 indicating identical clusters and 1 indicating maximally different clusters. Threshold refers to the distance threshold used to assign clusters with HDBSCAN.

Pathogen	Method	VI	Threshold
Influenza H3N2	t-SNE	0.04	2.0
	UMAP	0.08	1.0
	MDS	0.10	3.5
	PCA	0.19	0.5
SARS-CoV-2 (Nextstrain clade)	t-SNE	0.07	1.0
	MDS	0.15	0.0
	UMAP	0.16	0.5
	PCA	0.36	4.5
SARS-CoV-2 (Nextclade pango)	t-SNE	0.12	1.0
	MDS	0.23	0.0
	UMAP	0.25	0.5
	PCA	0.42	4.5

As expected, the clusters for each method generally corresponded to larger  
 506 phylogenetic clades (Fig 4). Clusters from t-SNE most accurately captured expert clade  
 507 assignments (normalized VI=0.04) with nine clusters. These clusters captured broader  
 508 phylogenetic clades (A1, A1b, A2, A3, A4, 3c2.A, and 3c3.A) but failed to distinguish  
 509 between A1b and its descendants. Clusters from UMAP performed nearly as well  
 510 (normalized VI=0.08) with six clusters. These clusters captured broader clades, but  
 511 they failed to distinguish among A1 and A1a, A1b and its subclades, and 3c2.A and A4.  
 512 The nine MDS clusters were more than twice as far from expert clades as t-SNE clusters  
 513 (normalized VI=0.10), but this difference in accuracy appeared to be driven primarily  
 514 by the cost of more samples that HDBSCAN failed to assign to clusters in the MDS  
 515

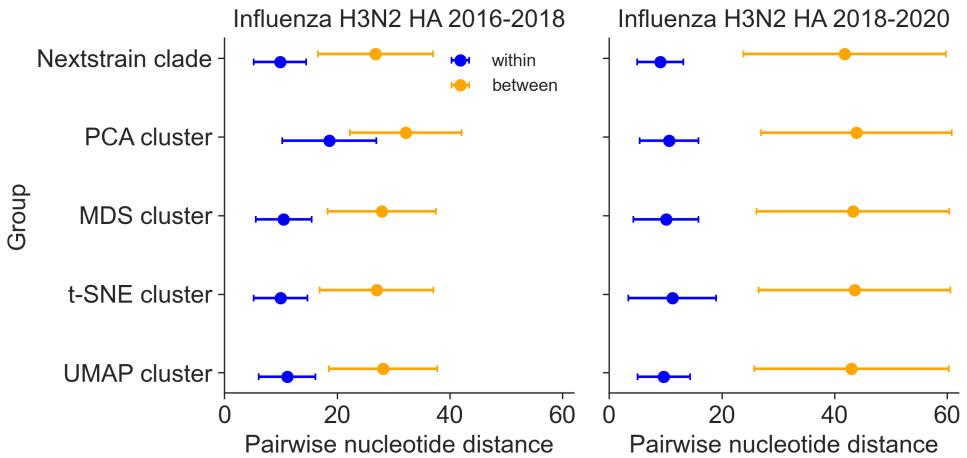
embedding. MDS clusters captured most of the larger clades (A1, A2, A3, A4, 3c2.A, and 3c3.A), but they also collected A1 and its descendants into two large clusters. The PCA embedding produced the lowest accuracy (normalized VI=0.19) and fewest clusters (N=3). PCA's three clusters corresponded to some of the most distantly-related and ancestral clades (3c2.A, 3c3.A, and A2). We identified 30 cluster-specific mutations for all three PCA clusters, 34 for seven of the nine MDS clusters, 32 for six of nine t-SNE clusters, and 25 for four of the six UMAP clusters (S1 Table). We also found that pairwise genetic distances between sequences in the same MDS, t-SNE, or UMAP clusters matched the genetic distances between sequences in the same Nextstrain clades (S4 Fig). 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 and optimal cluster parameters could effectively cluster previously unseen sequences, we applied each method to the late H3N2 HA dataset (2018–2020), clustered sequences in the embedding space with HDBSCAN using the optimal minimum distance threshold from the early dataset, and calculated the accuracy of the cluster assignments based on previously defined clades. Unlike the early H3N2 HA dataset, the late dataset contained fewer clades (S5 Fig) and a greater genetic distance between samples in clades (S4 Fig). Clusters from all four methods generally captured phylogenetic clades (Fig. 5 and S5 Fig). The MDS clusters most accurately captured expert clades (normalized VI=0.07) with six clusters corresponding to the largest clades (Fig. 5 and S6 Fig). MDS split A3 samples into two widely separated groups in its Euclidean space, indicating substantial within-clade genetic differences. On inspection of this clade, we found recurrent HA1 substitutions of 135K, 142G, and 193S in multiple subclades that MDS could not effectively represent. Clusters from t-SNE, UMAP, and PCA followed closely in accuracy (normalized VI=0.08, 0.08, and 0.09) with 5, 7, and 6 clusters, respectively. We identified 25 cluster-specific mutations for four of the six PCA clusters, 51 for all six MDS clusters, 53 for all five t-SNE clusters, and 30 for five of the seven UMAP clusters (S1 Table). Pairwise genetic distances within clusters generally matched the diversity within Nextstrain clades (S4 Fig). Cluster accuracies remained robust to sampling density and composition (S7 Fig). These results show that all four methods can produce



**Fig 4. Phylogenetic trees (left) and embeddings (right) of early (2016–2018) influenza H3N2 HA sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Nextstrain clades).**

well-supported clusters that accurately capture known genetic groups when applied to previously unseen H3N2 HA samples.

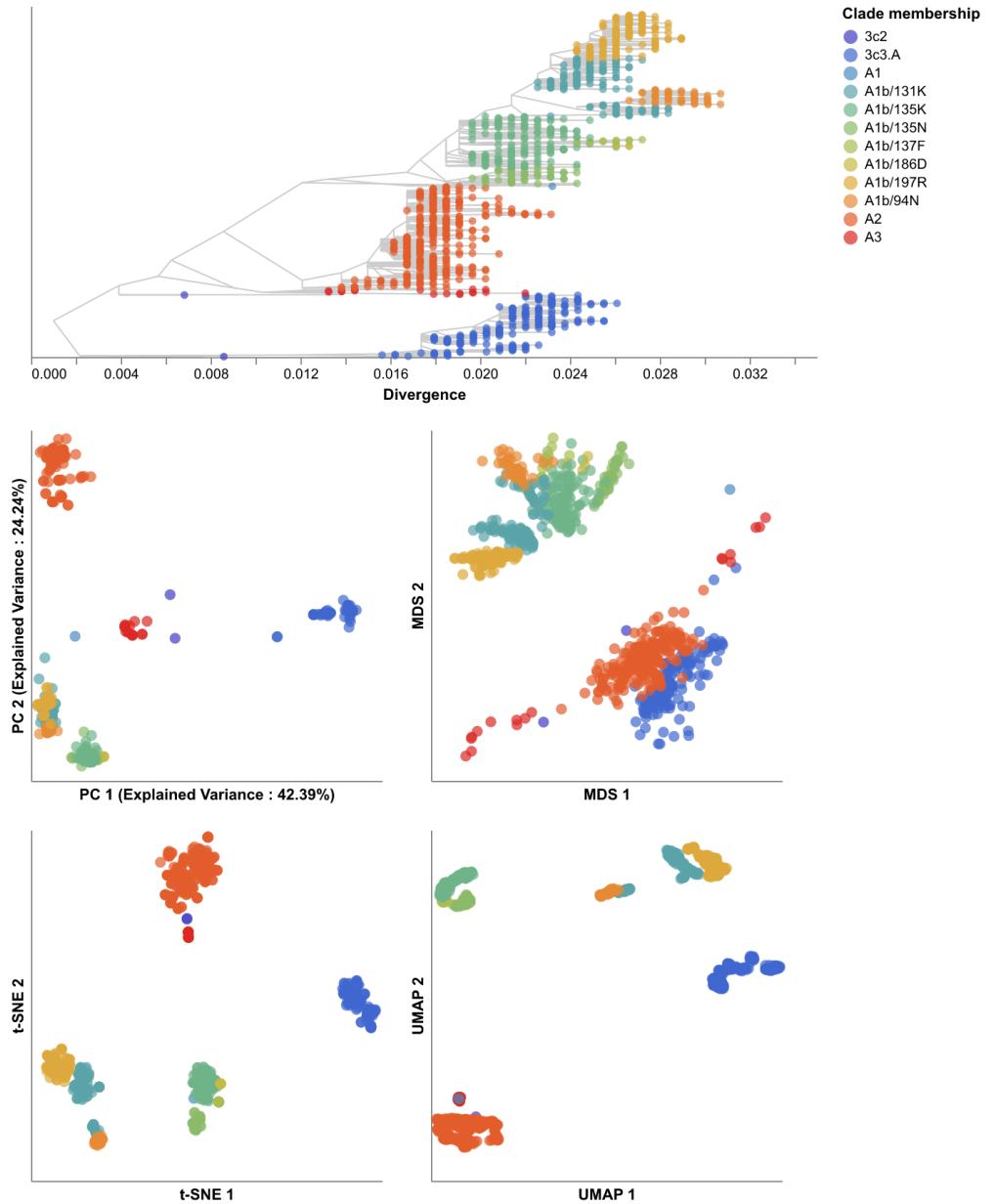


**S4 Fig. Pairwise nucleotide distances for early (2016–2018) and late (2018–2020) influenza H3N2 HA sequences within and between genetic groups defined by Nextstrain clades and clusters from PCA, MDS, t-SNE, and UMAP embeddings.**

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

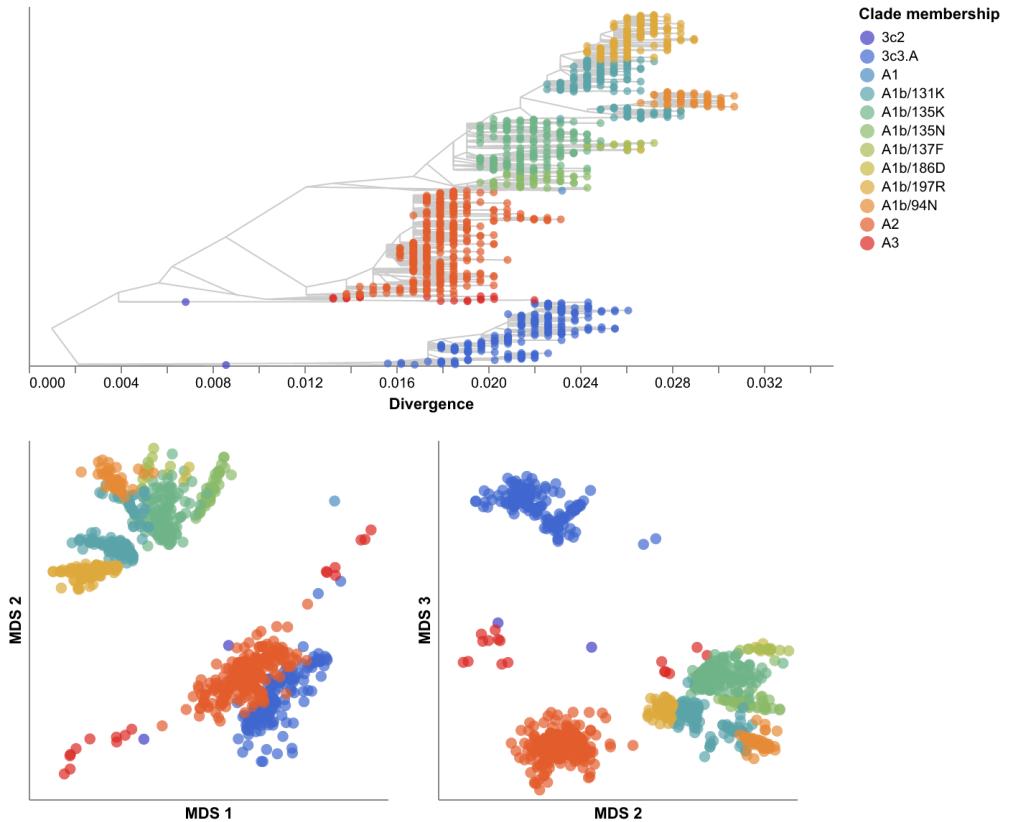
Given that clusters from embedding methods could recapitulate expert-defined clades, we measured how well the same methods could capture reassortment events between multiple gene segments as detected by biologically-informed computational models. Evolution of HA and NA surface proteins contributes to the ability of influenza viruses to escape existing immunity [57] and HA and NA genes frequently reassort [5, 6, 59]. Therefore, we focused our reassortment analysis on HA and NA sequences, sampling 1,607 viruses collected between January 2016 and January 2018 with sequences for both genes. We aligned these sequences to a common reference (A/Beijing/32/1992), inferred HA and NA phylogenies, and applied TreeKnit to both trees to identify maximally compatible clades (MCCs) that represent reassortment events [11]. Of the 191 reassortment events identified by TreeKnit, 15 (8%) contained at least 10 samples representing 1,062 samples (66%).

We created PCA, MDS, t-SNE, and UMAP embeddings from the HA alignments and from merged HA and NA alignments. We identified clusters in both HA-only and HA/NA embeddings and calculated the VI distance between these clusters and the MCCs identified by TreeKnit. We expected that clusters from HA-only embeddings



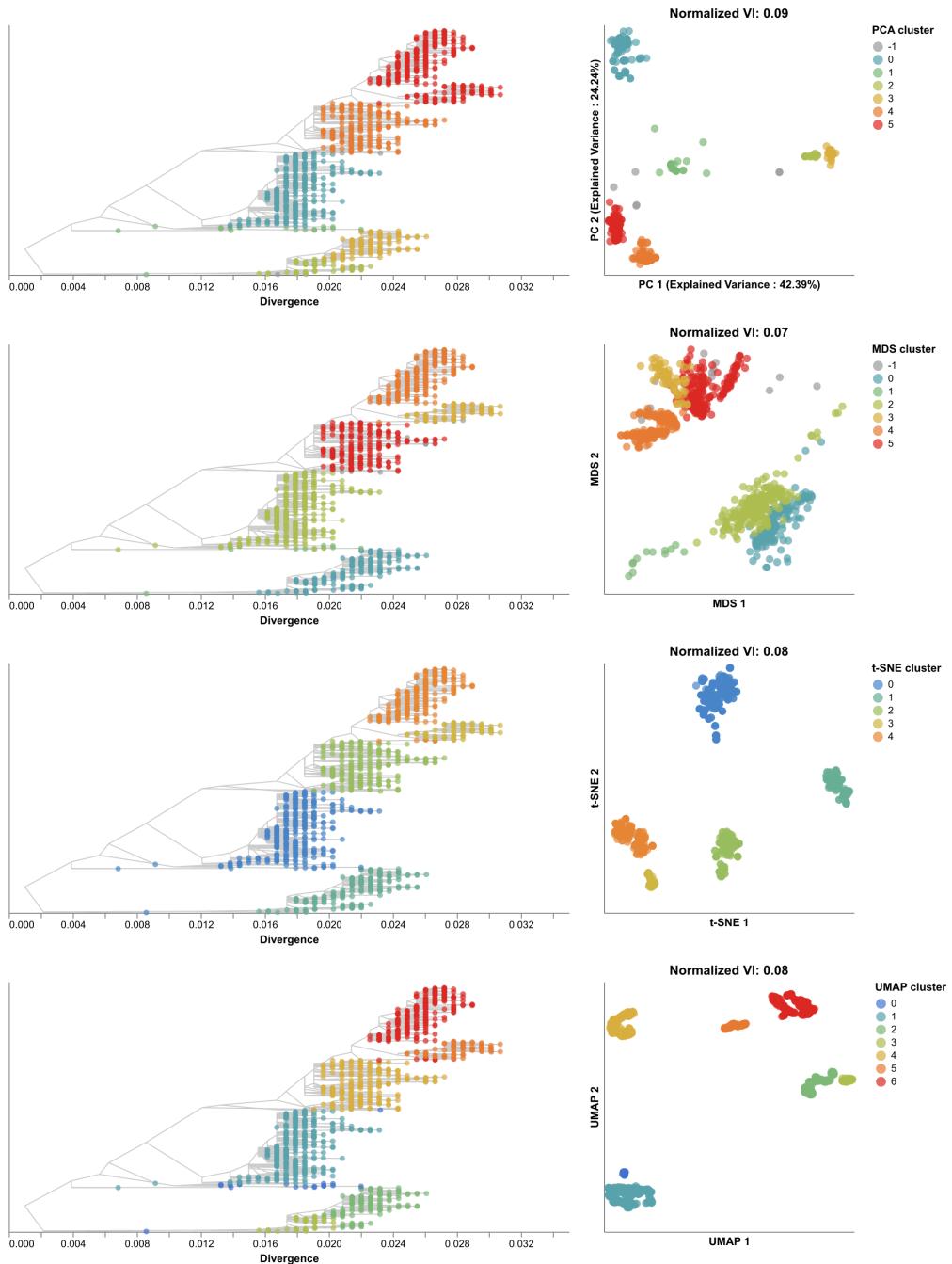
**S5 Fig. Phylogeny of early (2018–2020) influenza H3N2 HA sequences plotted by nucleotide substitutions per site on the x-axis (top) and low-dimensional embeddings of the same sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips in the tree and embeddings are colored by their Nextstrain clade assignment.**

could only reflect reassortment events when the HA clade involved in reassortment happened to carry characteristic nucleotide mutations. We expected that the VI distances for clusters from HA/NA embeddings would improve on the baseline distances calculated with the HA-only clusters.



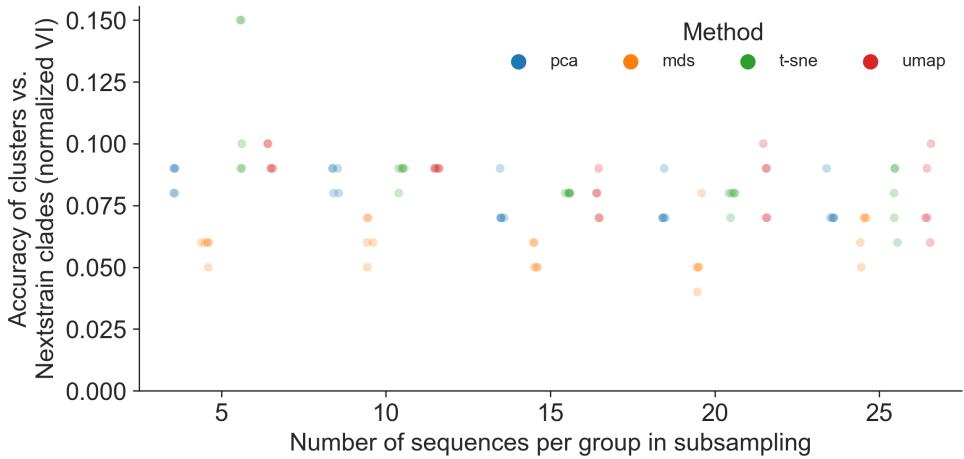
**S6 Fig. MDS embeddings for late (2018–2020) influenza H3N2 HA sequences showing all three components.**

All embedding methods produced more accurate clusters from the HA/NA alignments than the HA-only alignments (Fig. 6). HA/NA clusters from MDS reduced the distance to known reassortment events by 69% from a normalized VI value of 0.16 with HA only to 0.05. Adding NA to HA only modestly improved PCA, t-SNE, UMAP clusters, reducing distances by 0.05, 0.04, and 0.03, respectively. Embeddings with both genes also produced more clusters than the HA-only embeddings with one additional cluster in PCA (S9 Fig), 11 in MDS (S10 Fig), five in t-SNE (S11 Fig), and one in UMAP (S12 Fig). With the exception of PCA, all embeddings of HA/NA alignments produced distinct clusters for the known reassortment event within clade A2 [59] as represented by MCCs 14 and 11 (S8 Fig). Other larger events like those represented by MCCs 9 and 13 mapped far apart in all HA/NA embeddings except PCA. Smaller events like the reassortment represented by MCCs 10 and 6 only received separate clusters in the MDS embedding of HA and NA (S10 Fig). We noted that some of the additional clusters in HA/NA embeddings likely also reflected genetic diversity in NA



**Fig 5. Phylogenetic trees (left) and embeddings (right) of late (2018–2020) H3N2 HA sequences colored by HDBSCAN cluster.** Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Nextstrain clades).

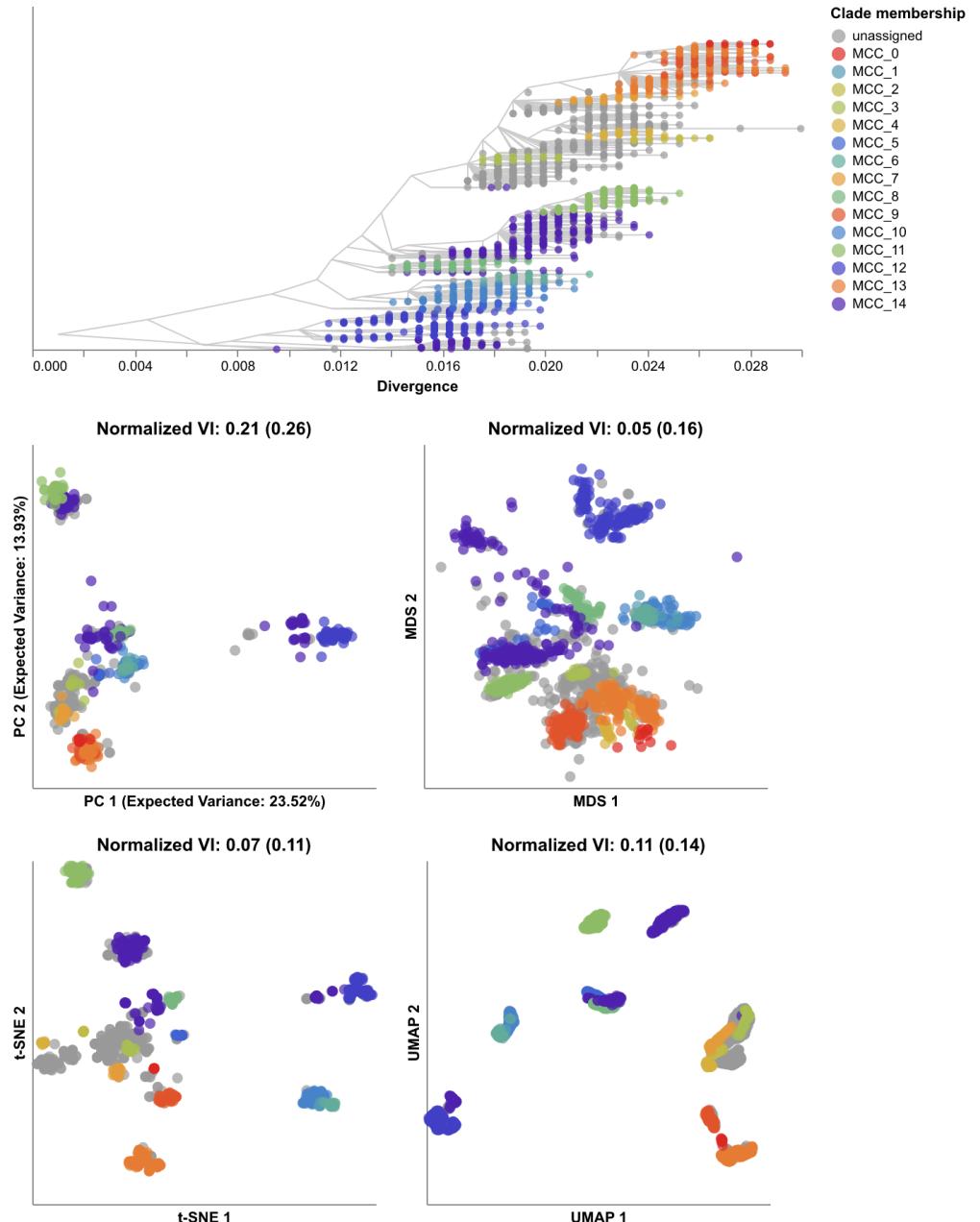
that was independent of reassortment between HA and NA. These results suggest that a single embedding of multiple gene segments could identify biologically meaningful clusters within and between all genes.



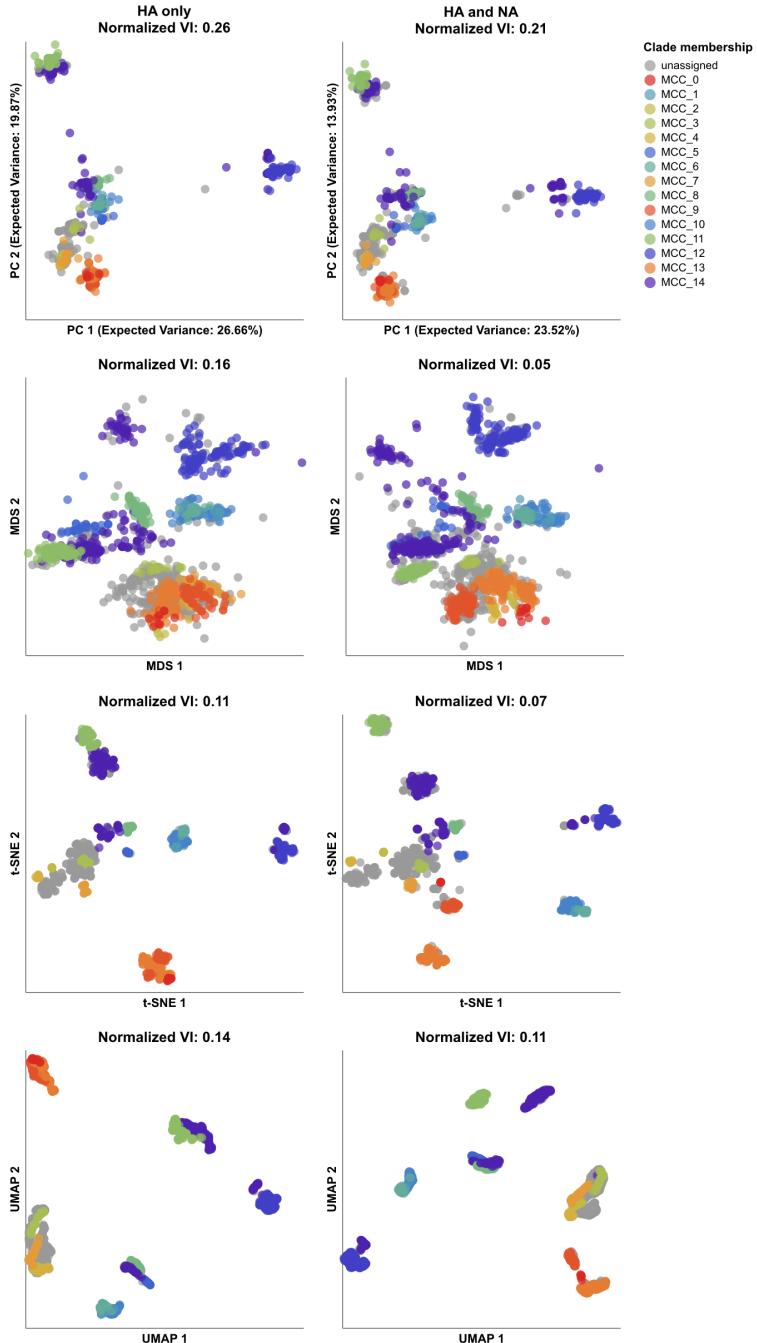
**S7 Fig. Replication of cluster accuracy per embedding method for late (2018–2020) influenza H3N2 HA sequences across different sequences per group sampled from the original dataset and five replicates per sampling density.**

### SARS-CoV-2 clusters recapitulate broad genetic groups corresponding to Nextstrain clades

SARS-CoV-2 poses a greater challenge to embedding methods than seasonal influenza, with an unsegmented genome an order of magnitude longer than influenza's HA or NA [60], a mutation rate in the spike surface protein subunit S1 that is four times higher than influenza H3N2's HA rate [61], and increasingly common recombination [62, 63]. However, multiple expert- and model-based clade definitions exist for SARS-CoV-2, enabling comparison between clusters from embeddings and known genetic groups. These definitions span from broad genetic groups named by the WHO as “variants of concern” (e.g., “Alpha”, “Beta”, etc.) [64] or systematically defined by the Nextstrain team [51–53] to smaller, emerging genetic clusters defined by Pangolin [17]. As with seasonal influenza, we defined an early SARS-CoV-2 dataset spanning from January 2020 to January 2022, embedded genomes with the same four methods, and identified HDBSCAN clustering parameters that minimized the VI distance between embedding clusters and previously defined genetic groups as defined by Nextstrain clades and collapsed “Nextclade pango” lineages (see Methods). Using these optimal cluster parameters, we produced clusters from embeddings of a late SARS-CoV-2 dataset spanning from January 2022 to November 2023 and calculated the



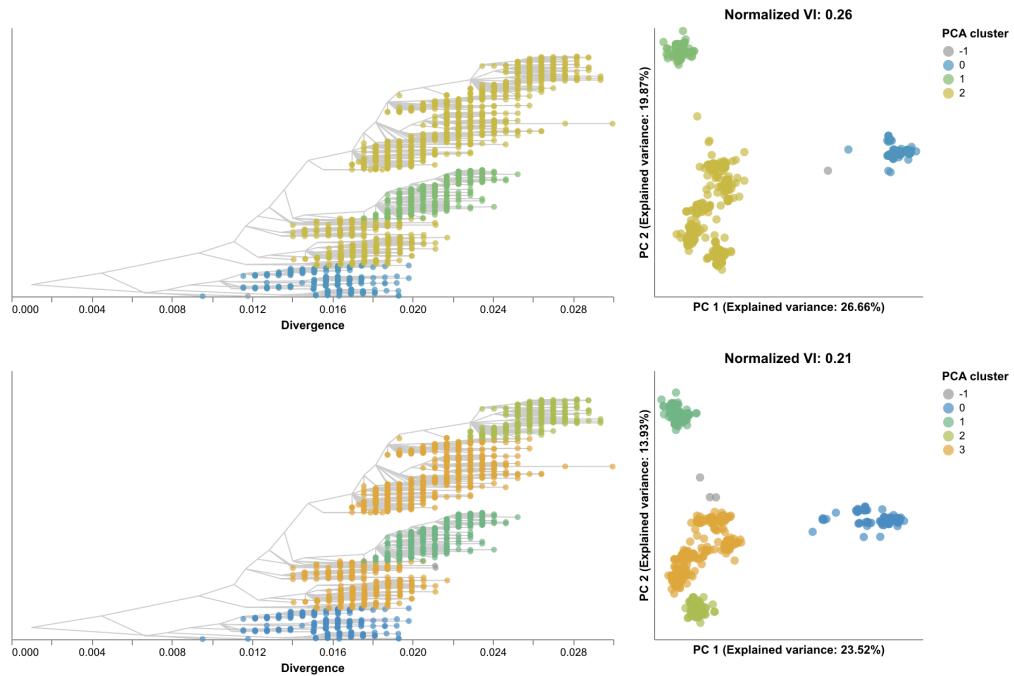
**Fig 6.** Phylogeny of early (2016–2018) influenza H3N2 HA sequences plotted by nucleotide substitutions per site on the x-axis (top) and low-dimensional embeddings of the same HA sequences concatenated with matching NA sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips in the tree and embeddings are colored by their TreeKnit Maximally Compatible Clades (MCCs) label which represents putative HA/NA reassortment groups. The first normalized VI values per embedding reflect the distance between HA/NA clusters and known genetic groups (MCCs). VI values in parentheses reflect the distance between HA-only clusters and known genetic groups.



**S8 Fig.** Embeddings influenza H3N2 HA-only (left) and combined HA/NA (right) showing the effects of additional NA genetic information on the placement of reassortment events detected by TreeKnit (MCCs).

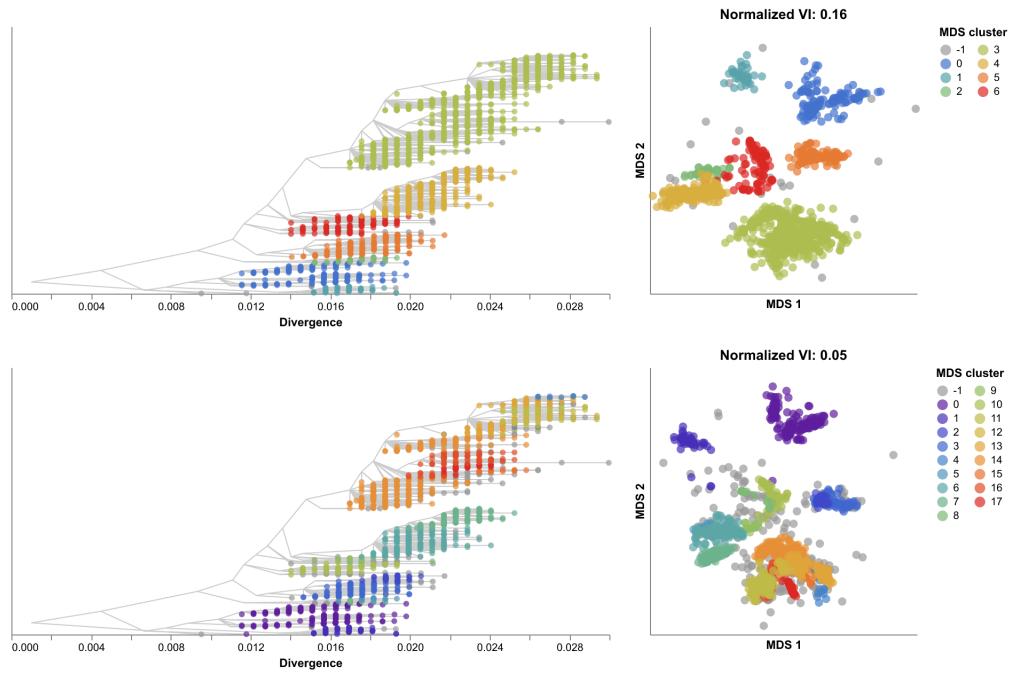
VI distance between those clusters and known genetic groups.

The early SARS-CoV-2 dataset represented 24 Nextstrain clades and 35 collapsed Nextclade pango lineages. With the exception of PCA, all other embedding methods



**S9 Fig.** PCA embeddings for influenza H3N2 HA sequences only (top row) and HA/NA sequences combined (bottom row) showing the HA trees colored by clusters identified in each embedding (left) and the corresponding embeddings colored by cluster (right).

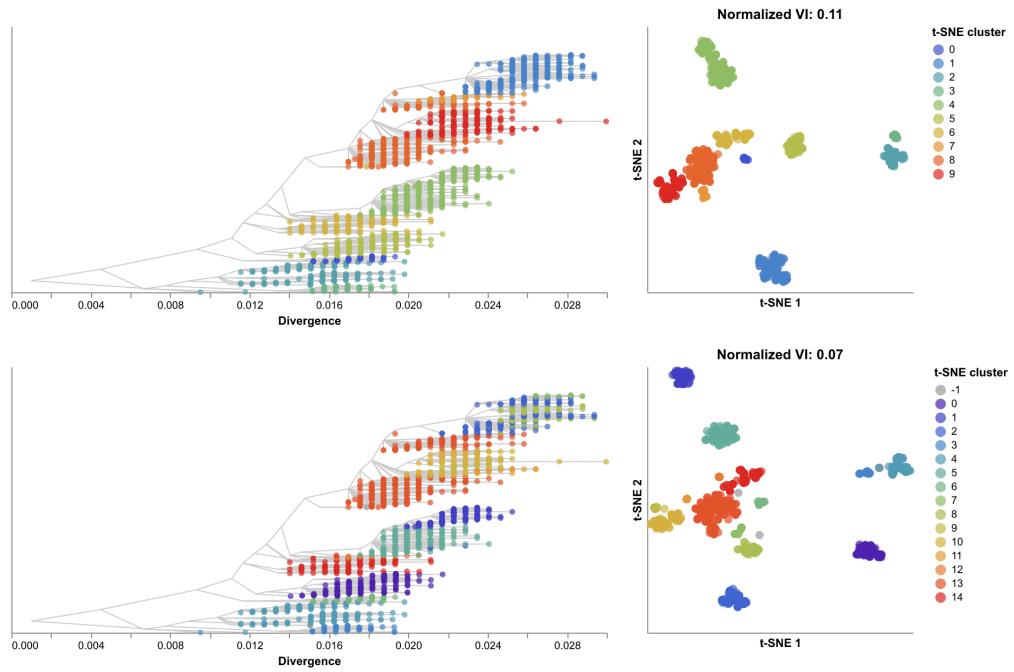
placed samples from the same Nextstrain clades closer together and closely related  
Nextstrain clades near each other (Fig. 7). For example, the most genetically distinct  
clades like 21J (Delta) and 21K (Omicron) placed farthest from other clades, while all  
Delta clades (21A, 21I, and 21J) placed close together (Fig. 7, S13 Fig). As we saw with  
embeddings of H3N2 HA sequences, MDS placed related clades closer together on a  
continuous scale, while t-SNE and UMAP produced more clearly separate groups of  
samples. Unlike the H3N2 HA analysis, the PCA embedding of SARS-CoV-2 sequences  
failed to create any genetically meaningful clusters. We suspected that PCA  
components reflected variation in missing (“N”) or gap (“-”) characters that we  
represented with a separate character state than the standard nucleotide characters of  
A, C, G, and T. We plotted the PC1 value of each sample against the number of  
missing bases in its alignment and confirmed that missing data explained a substantial  
proportion of variation in PC1 but not all (S14 Fig, Pearson’s  $R^2 = 0.304$ ). When we  
compared embedding clusters to Nextclade pango lineages, we did not observe the same  
clear grouping as we did with Nextstrain clades. For example, the Nextstrain clade 21J



**S10 Fig. MDS embeddings for influenza H3N2 HA sequences only (top row) and HA/NA sequences combined (bottom row) showing the HA trees colored by clusters identified in each embedding (left) and the corresponding embeddings colored by cluster (right).**

(Delta) contained 11 pango lineages that all appeared to map into the same overlapping space in MDS, t-SNE, and UMAP embeddings (S15 Fig). These results suggest that distance-based embedding methods can recapitulate broader genetic groups of SARS-CoV-2, but that these methods lack the resolution of finer groups defined by Pangolin.

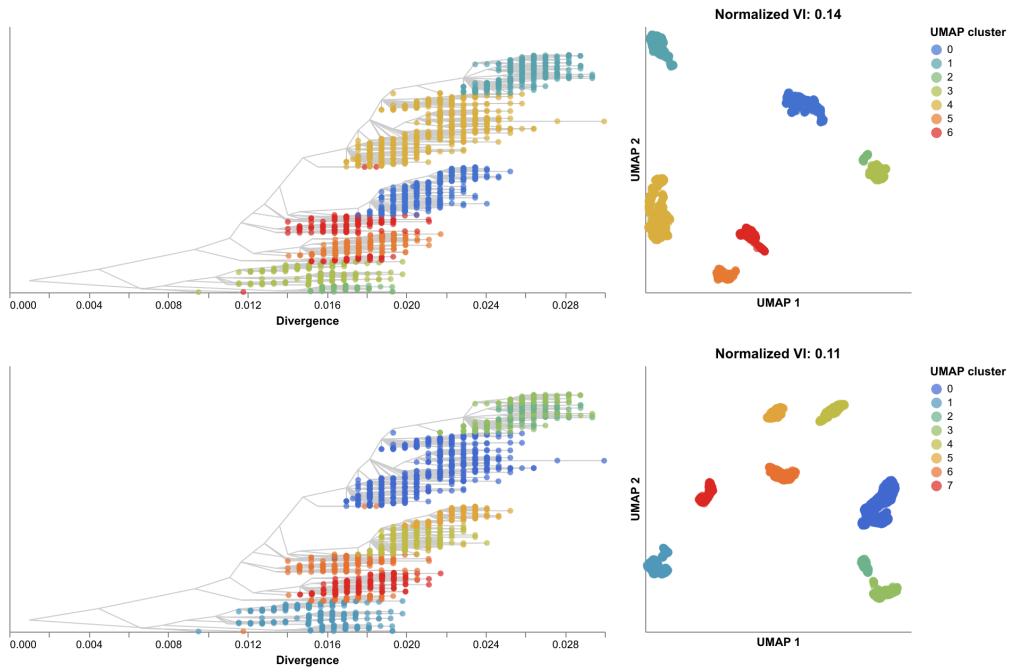
We quantified the maintenance of local and global structure in early SARS-CoV-2 embeddings by fitting a linear model between pairwise genetic and Euclidean distances of samples. As we expected from the qualitative evaluation of the PCA embedding above, we found no relationship between Euclidean distance in PCA and genetic distance in alignments (Fig. 8). In contrast, the MDS embedding produced a strong linear mapping across the range of observed genetic distances (Pearson's  $R^2 = 0.92$ ). Both t-SNE and UMAP maintained intermediate degrees of linearity (Pearson's  $R^2 = 0.63$  and  $R^2 = 0.55$ , respectively). These embeddings placed the most genetically similar samples close together and the most genetically distinct farther apart. However, these embeddings did not consistently place pairs of samples with intermediate genetic distances at an intermediate distance in Euclidean space. The linear relationship for



**S11 Fig.** t-SNE embeddings for influenza H3N2 HA sequences only (top row) and HA/NA sequences combined (bottom row) showing the HA trees colored by clusters identified in each embedding (left) and the corresponding embeddings colored by cluster (right).

genetically similar samples in t-SNE remained consistent up to a genetic distance of approximately 30 nucleotides. The corresponding relationship for UMAP only remained consistent up to a genetic distance of approximately 10 nucleotides.

We identified clusters in embeddings from early SARS-CoV-2 data using cluster parameters that minimized the normalized VI distance between clusters and known genetic groups. Since Nextstrain clades and Nextclade pango lineages represented different resolutions of genetic diversity, we identified separate optimal parameters for clusters compared to each of these known genetic groups. When comparing clusters to Nextstrain clades, the t-SNE embedding produced the most accurate clusters with a normalized VI of 0.07 (N=19 clusters, minimum distance of 1.0) (Fig. 9, Table 1). MDS and UMAP produced similarly accurate clusters with normalized VIs of 0.15 (N=16) and 0.16 (N=6) at minimum distances of 0 and 0.5, respectively. As expected, PCA produced the least accurate clusters with a normalized VI of 0.36 (N=2, minimum distance of 4.5). We found 18 cluster-specific mutations for one of the two PCA clusters, 232 for nine of 16 MDS clusters, 363 for 15 of 19 t-SNE clusters, and 155 for five of six UMAP clusters (S1 Table). Clusters from t-SNE produced within-group genetic

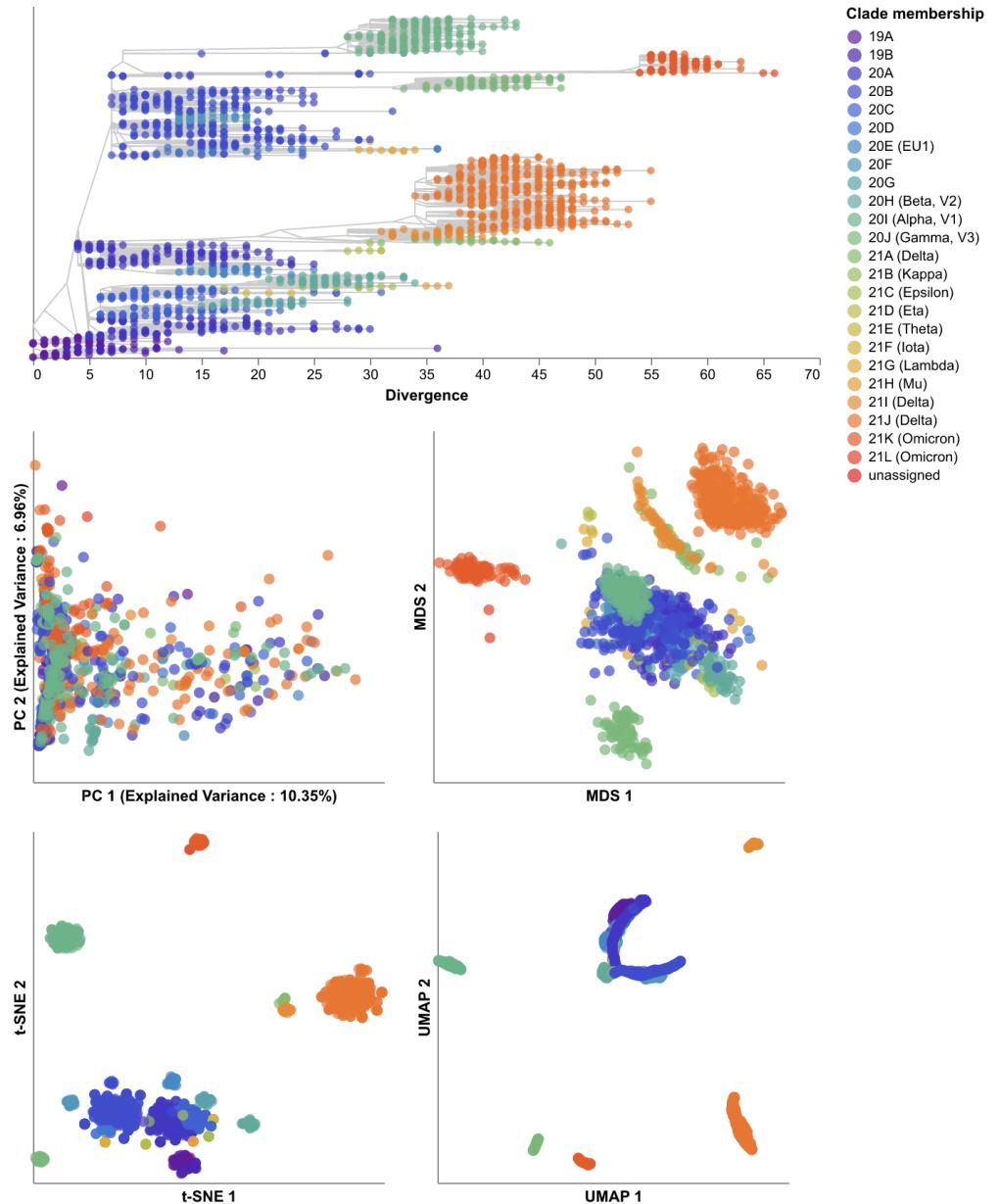


**S12 Fig.** UMAP embeddings for influenza H3N2 HA sequences only (top row) and HA/NA sequences combined (bottom row) showing the HA trees colored by clusters identified in each embedding (left) and the corresponding embeddings colored by cluster (right).

distances that were most similar to distances within Nextstrain clades (S16 Fig).

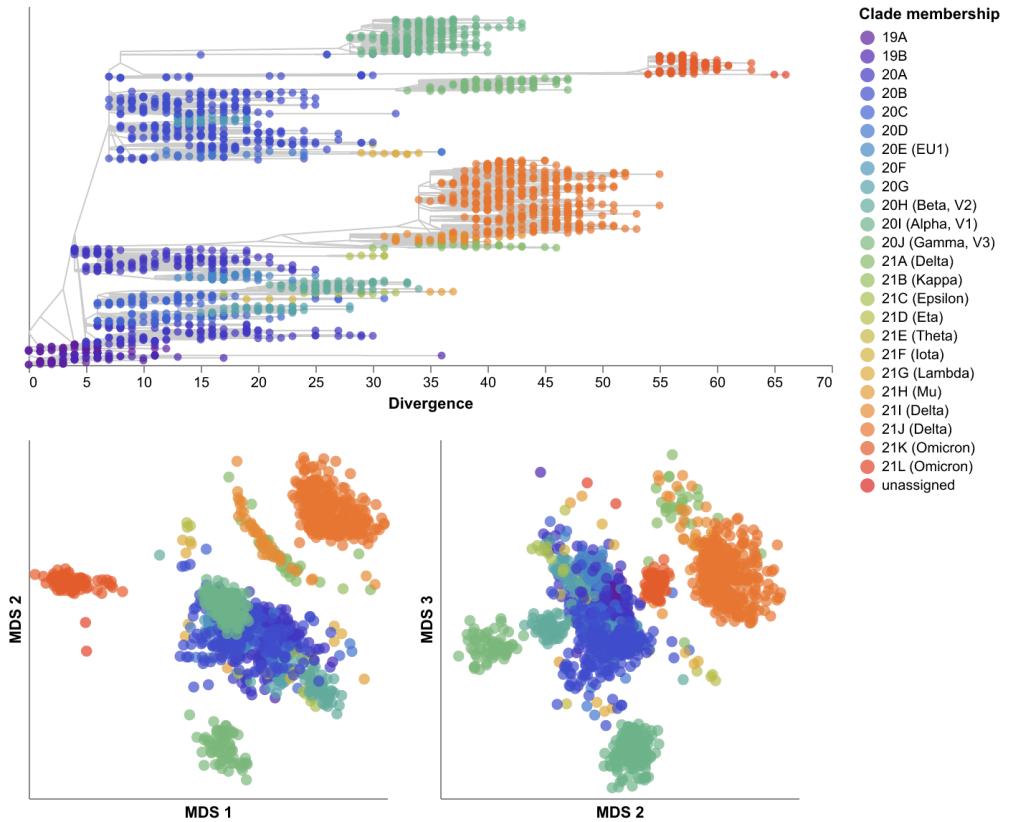
When comparing clusters to Nextclade pango lineages, all four methods produced less accurate clusters (S17 Fig). Clusters from t-SNE were the most accurate with a VI of 0.12. MDS and UMAP clusters performed similarly with VIs of 0.23 and 0.25. PCA clusters remained the least accurate with a VI of 0.42. The optimal minimum distances for all four methods remained the same with Nextclade pango lineages as when trained with Nextstrain clades. These results confirm quantitatively that these embeddings methods can accurately capture broader genetic diversity of SARS-CoV-2, but most methods cannot distinguish between fine resolution genetic groups identified by Pangolin. However, we observed greater pairwise genetic distances within collapsed Nextclade pango lineages than within Nextstrain clades, suggesting that Pangolin lineages were not as tightly scoped as we originally expected (S16 Fig).

To test the optimal cluster parameters identified above, we applied embedding methods to late SARS-CoV-2 data and compared clusters from these embeddings to known genetic groups. Of the 17 Nextstrain clades defined during this time period, 14 (82%) descended from Omicron and represented 1,495 (90%) of all samples in the



**Fig 7.** Phylogeny of early (2020–2022) SARS-CoV-2 sequences plotted by number of nucleotide substitutions from the most recent common ancestor on the x-axis (top) and low-dimensional embeddings of the same sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips in the tree and embeddings are colored by their Nextstrain clade assignment.

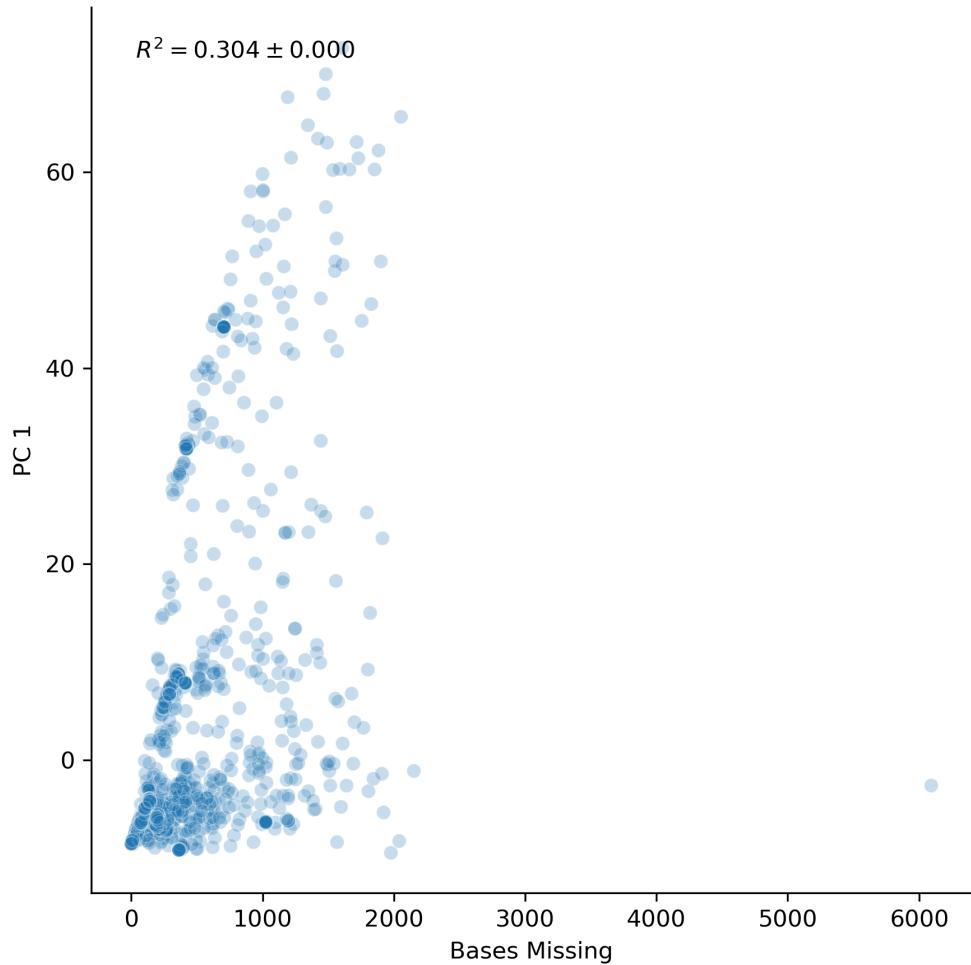
dataset. Of the 51 Nextclade pango lineages, 20 originated from a recombination event and corresponded to 521 (31%) of all samples. The clusters from embeddings of these more recent SARS-CoV-2 sequences performed as well or better than the clusters from earlier SARS-CoV-2 sequences (Fig. 10). Clusters from t-SNE most accurately matched



**S13 Fig. MDS embeddings for early SARS-CoV-2 sequences showing all three components.**

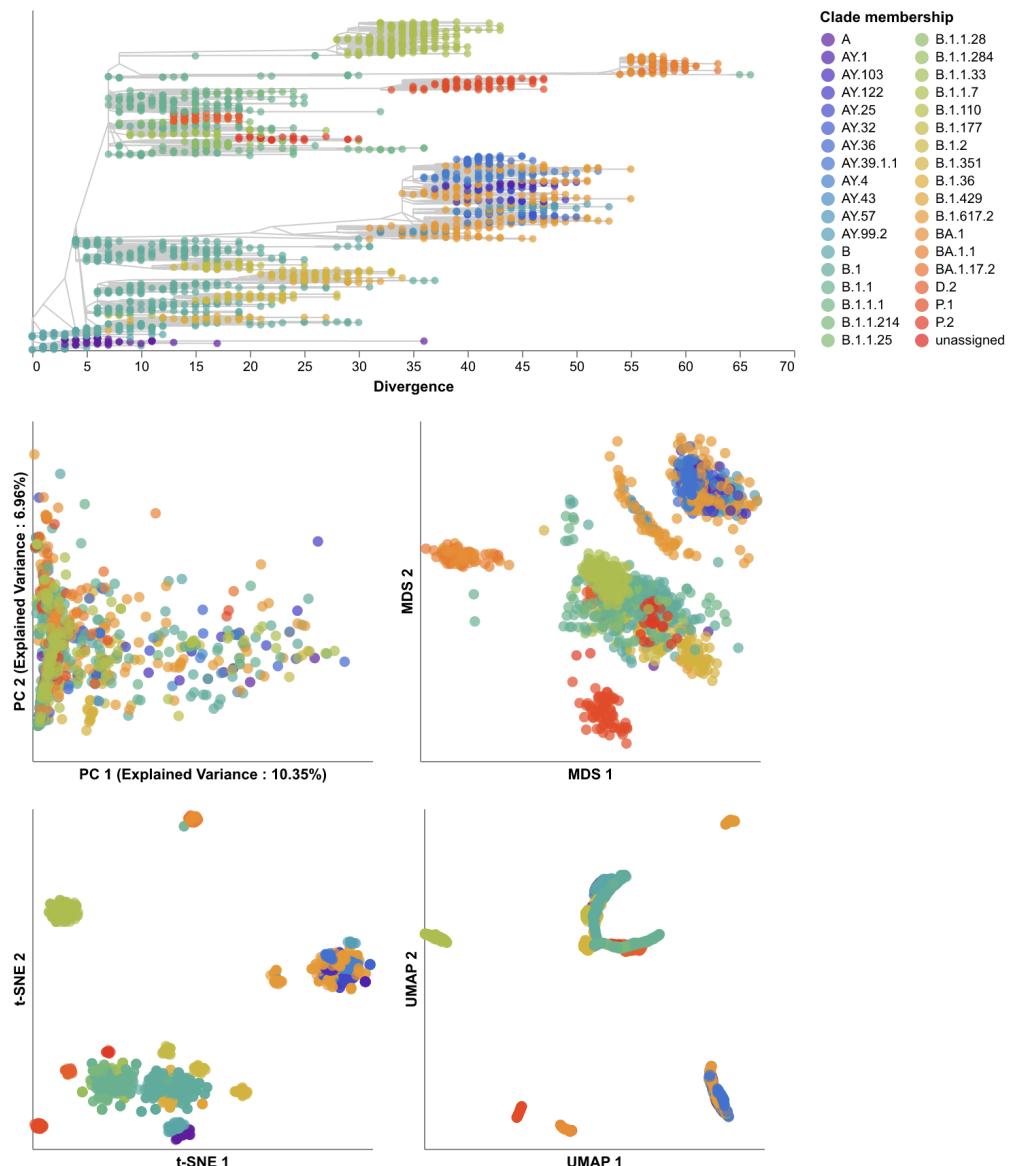
Nextstrain clades (normalized VI=0.08) with 22 clusters. Clusters from UMAP followed (normalized VI=0.13) with nine clusters and MDS produced 10 clusters (normalized VI=0.15). We found 50 cluster-specific mutations for two of three PCA clusters, 75 for nine of 10 MDS clusters, 93 for 16 of 22 t-SNE clusters, and 112 for eight of nine UMAP clusters (S1 Table). Clusters from t-SNE remained the most consistently accurate across different random samples of sequences from the same time period and different total samples (S18 Fig). UMAP clusters became more accurate and less variable as we added more sequences per group to the embeddings.

All methods produced less accurate representations of Nextclade pango lineages (S19 Fig). Clusters from t-SNE were twice as far from Nextclade pango lineages than Nextstrain clades (normalized VI=0.16). UMAP's clusters were nearly two times farther from pango lineages than Nextstrain clades (normalized VI=0.23). Clusters from MDS were 1.6 times as far from pango lineages as Nextstrain clades (normalized VI=0.24). These results replicate the patterns we observed with early SARS-CoV-2 data where



**S14 Fig.** Principal component 1 (PC1) of the PCA embedding for early SARS-CoV-2 data plotted by the number of missing (“N”) or gap (“-”) characters in the corresponding sample’s aligned sequence. Pearson’s  $R^2$  estimates the variation in PC1 explained by missing data.

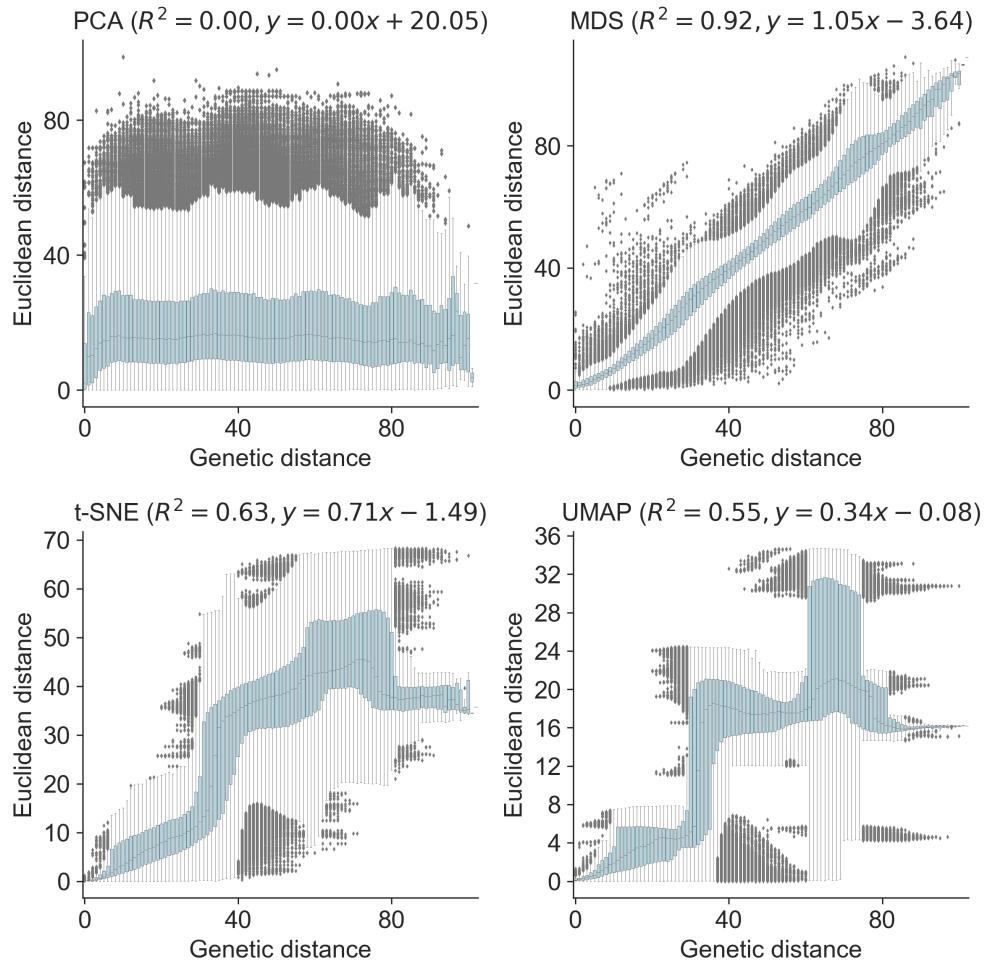
clusters from embeddings more effectively represented broader genetic diversity than the finer resolution diversity labeled by Pangolin. Unlike the Nextclade pango lineages in the early SARS-CoV-2 data, the lineages from the later data exhibited fewer pairwise genetic distances between samples in each lineage than samples in Nextstrain clades or any embedding cluster (S16 Fig).



**S15 Fig.** Phylogeny of early (2020–2022) SARS-CoV-2 sequences plotted by number of nucleotide substitutions from the most recent common ancestor on the x-axis (top) and low-dimensional embeddings of the same sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips in the tree and embeddings are colored by their collapsed Nextclade pango lineage assignment.

## Distance-based embeddings reflect SARS-CoV-2 recombination events

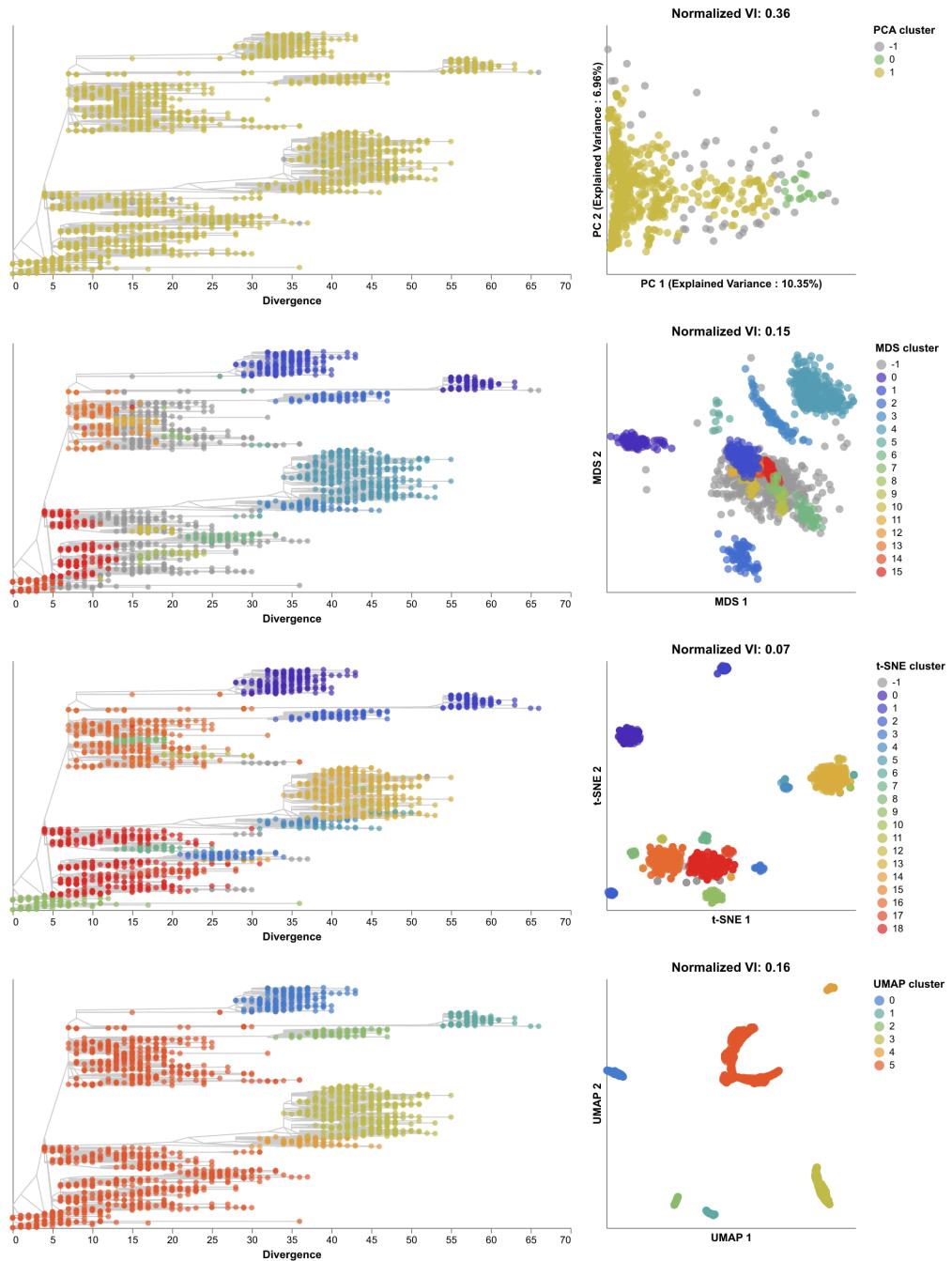
Finally, we tested the ability of sequence embeddings to capture patterns of recombination between known parental lineages of SARS-CoV-2. We reasoned that each



**Fig 8. Relationship between pairwise genetic and Euclidean distances in embeddings for early (2020–2022) SARS-CoV-2 sequences by PCA (upper left), MDS (upper right), t-SNE (lower left), and UMAP (lower right). Each boxplot represents the distribution of pairwise Euclidean distances at a given genetic distance.**

recombinant lineage,  $X$ , should always place closer to its parental lineages  $A$  and  $B$  than the parental lineages place to each other. Based on this logic, we calculated the average Euclidean distance between pairs of samples in lineages  $A$  and  $B$ ,  $A$  and  $X$ , and  $B$  and  $X$  for each embedding method (see Methods). We identified recombinant lineages that mapped closer to both of their parental lineages and those that mapped closer to at least one of the parental lineages.

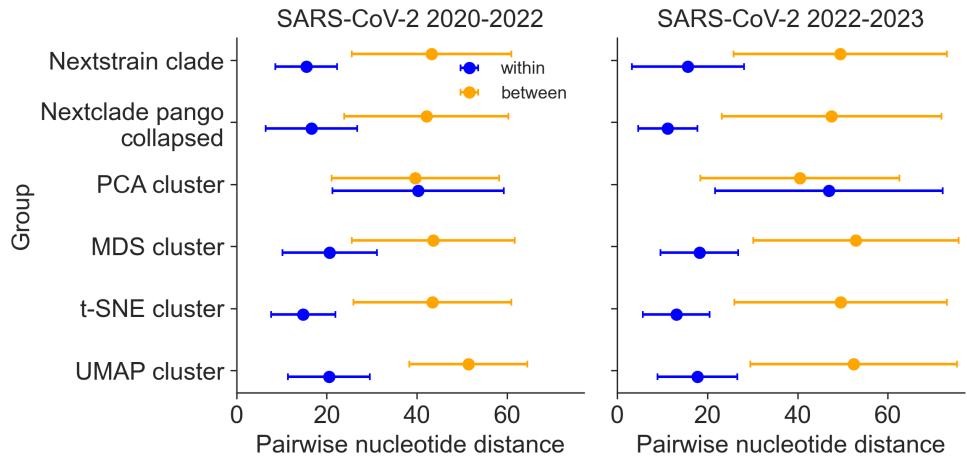
Seven of the ten recombinant lineages that we inspected had enough samples in both parental and recombinant lineages to calculate average pairwise distances (XD, XE, XF, XG, XBB, XBF, and XBL). MDS embeddings consistently placed recombinant lineages



**Fig 9. Phylogenetic trees (left) and embeddings (right) of early (2020–2022) SARS-CoV-2 sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Nextstrain clades).**

between the parental lineages in all seven cases (S2 Table). Embeddings from t-SNE and UMAP performed nearly as well, placing five of seven recombinant lineages between 709 parents. PCA performed the worst with only three recombinant lineages of seven 710

711

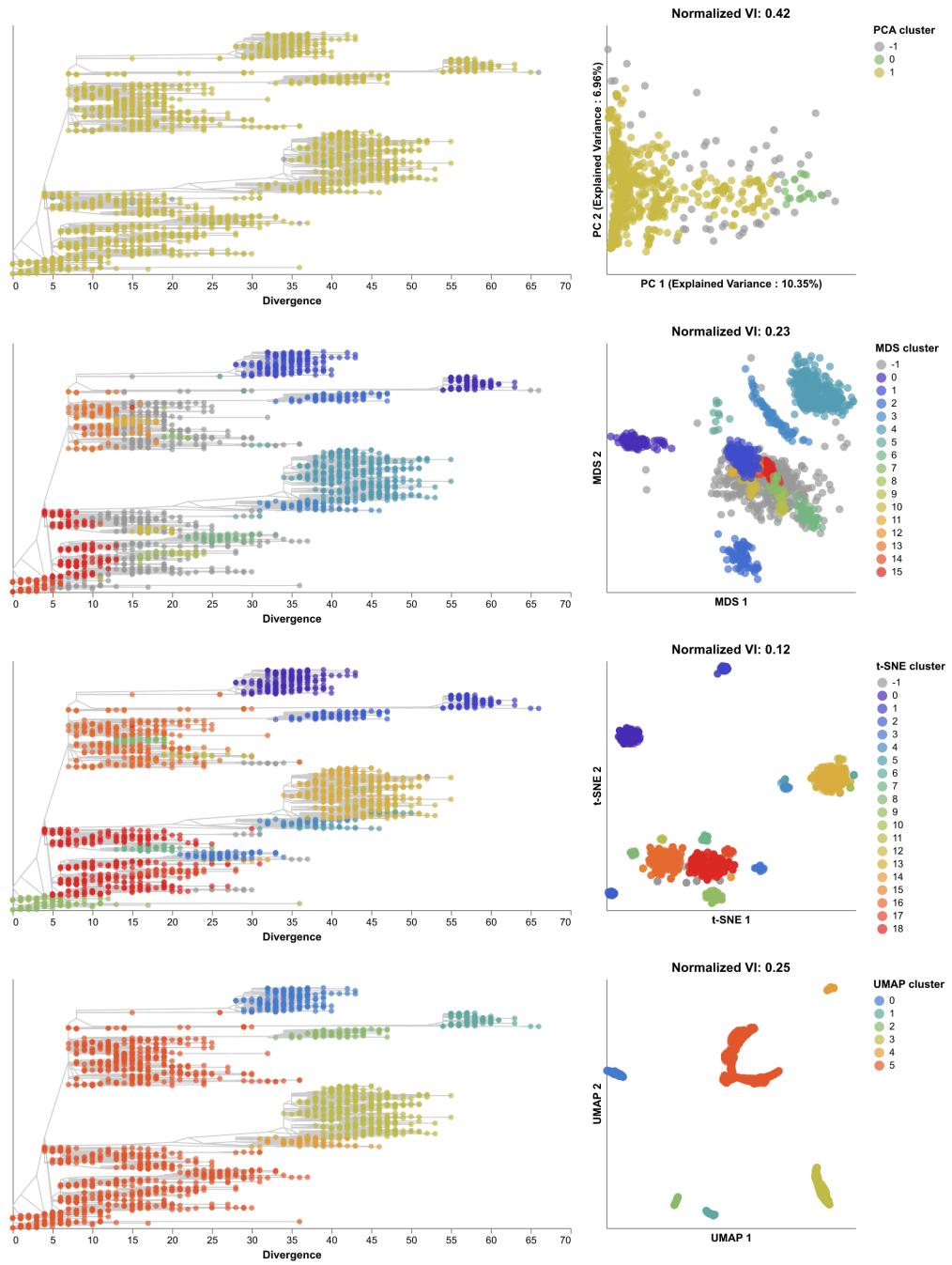


**S16 Fig. Pairwise nucleotide distances for early (2020-2022) and late (2022-2023) SARS-CoV-2 sequences within and between genetic groups defined by Nextstrain clades, collapsed Nextclade pango lineages, and clusters from PCA, MDS, t-SNE, and UMAP embeddings.**

placed between parentals. However, all recombinant lineages mapped closer to at least one parent in all embeddings. 712  
713

**S2 Table. Average Euclidean distances between each known recombinant,  $X$ , and its parental lineages  $A$  and  $B$  per embedding method.** Distances include average pairwise comparisons between  $A$  and  $B$ ,  $A$  and  $X$ , and  $B$  and  $X$ . Additional columns indicate whether each recombinant lineage maps closer to both parental lineages (or at least one) than those parents map to each other.

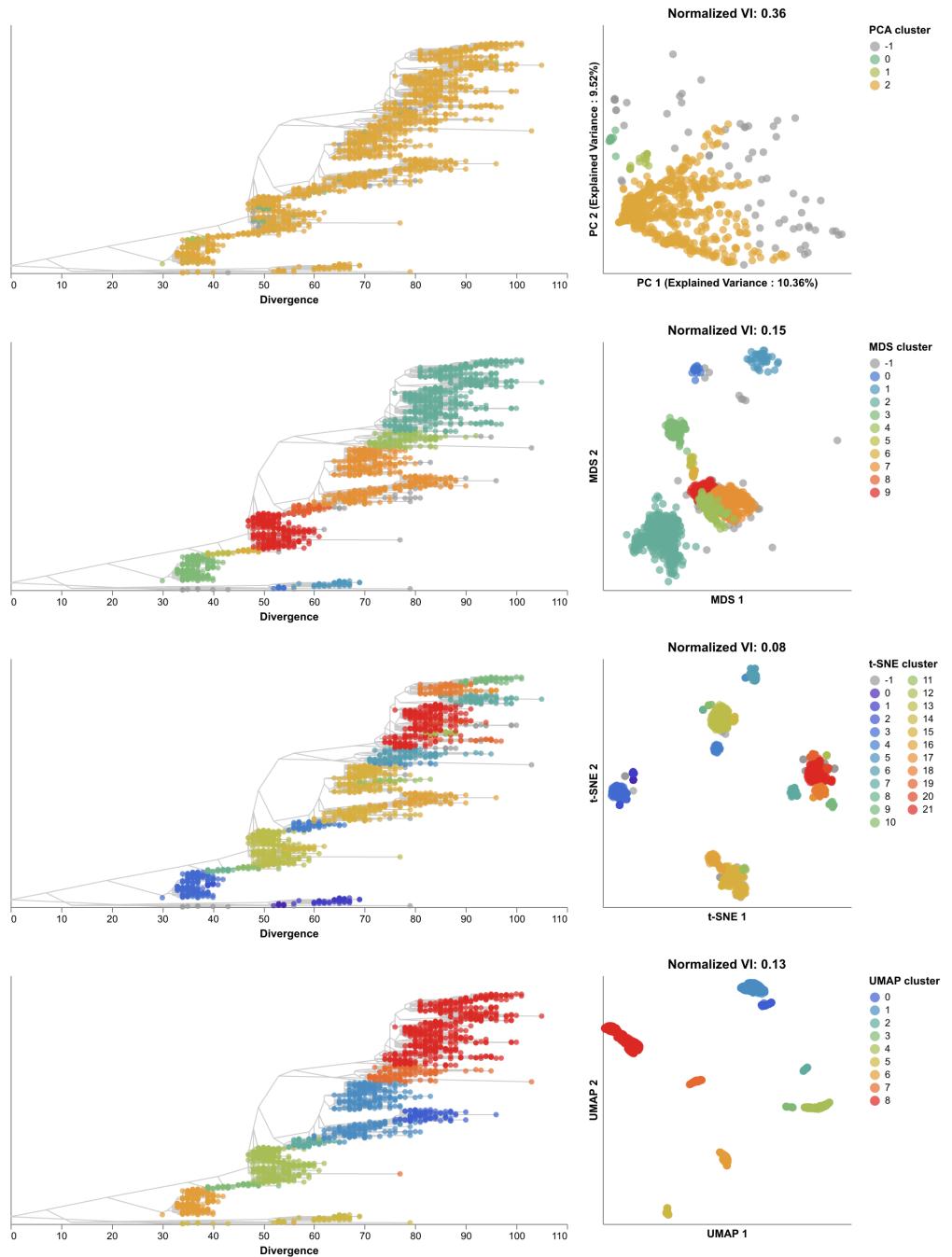
parental_A	parental_B	recombinant_X	method	distance_A,B	distance_A,X	distance_B,X	X_maps_closer_to_both_parents	X_maps_closer_to_any_parental
AY.4	BA.1	XD	PCA	14.14	7.51	15.17	False	True
BA.1	BA.2	XE	PCA	28.30	15.02	24.33	True	True
AY.4	BA.1	XF	PCA	14.14	9.16	16.51	False	True
BA.1	BA.2	XG	PCA	28.30	13.41	23.30	True	True
BJ.1	BM.1.1.1	XBB	PCA	23.28	17.44	24.87	False	True
BA.5.2.3	CJ.1	XBF	PCA	18.72	17.55	20.27	False	True
XBB.1	BA.2.75	XBL	PCA	22.95	20.43	20.52	True	True
AY.4	BA.1	XD	MDS	78.98	43.29	50.74	True	True
BA.1	BA.2	XE	MDS	48.67	32.41	21.46	True	True
AY.4	BA.1	XF	MDS	78.98	74.38	6.72	True	True
BA.1	BA.2	XG	MDS	48.67	39.75	13.91	True	True
BJ.1	BM.1.1.1	XBB	MDS	35.98	20.30	28.46	True	True
BA.5.2.3	CJ.1	XBF	MDS	43.51	39.39	9.68	True	True
XBB.1	BA.2.75	XBL	MDS	32.50	12.82	30.95	True	True
AY.4	BA.1	XD	t-SNE	6.65	1.66	5.35	True	True
BA.1	BA.2	XE	t-SNE	37.63	34.31	6.30	True	True
AY.4	BA.1	XF	t-SNE	6.65	9.33	2.83	False	True
BA.1	BA.2	XG	t-SNE	37.63	36.46	5.34	True	True
BJ.1	BM.1.1.1	XBB	t-SNE	27.56	4.87	31.21	False	True
BA.5.2.3	CJ.1	XBF	t-SNE	56.13	55.38	1.52	True	True
XBB.1	BA.2.75	XBL	t-SNE	31.98	4.21	31.26	True	True
AY.4	BA.1	XD	UMAP	12.19	0.88	11.54	True	True
BA.1	BA.2	XE	UMAP	15.83	11.93	4.83	True	True
AY.4	BA.1	XF	UMAP	12.19	11.80	0.77	True	True
BA.1	BA.2	XG	UMAP	15.83	12.48	4.02	True	True
BJ.1	BM.1.1.1	XBB	UMAP	13.71	1.33	14.76	False	True
BA.5.2.3	CJ.1	XBF	UMAP	17.82	17.87	0.38	False	True
XBB.1	BA.2.75	XBL	UMAP	14.81	1.75	13.46	True	True



**S17 Fig.** Phylogenetic trees (left) and embeddings (right) of early (2020–2022) SARS-CoV-2 sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (collapsed Nextclade pango lineages).

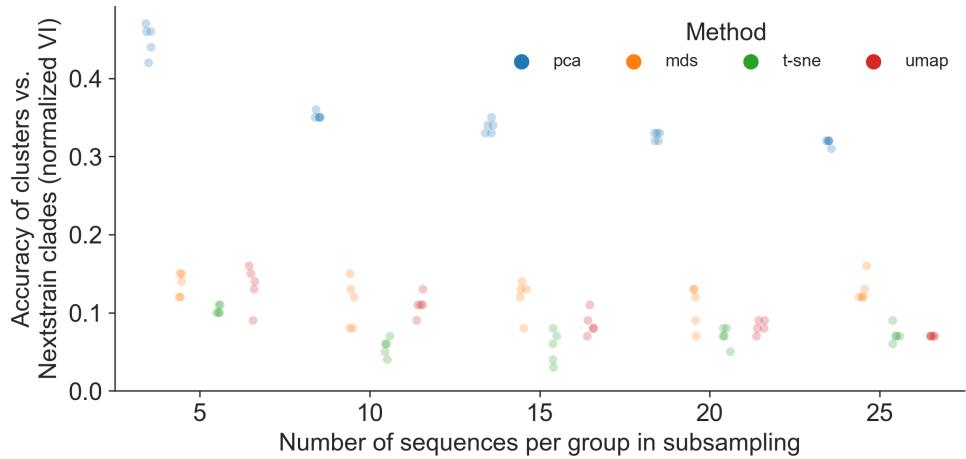
## Discussion

We applied four standard dimensionality reduction methods to simulated and natural genome sequences of two relevant human pathogenic viruses and found that the



**Fig 10. Phylogenetic trees (left) and embeddings (right) of late (2022–2023) SARS-CoV-2 sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Nextstrain clades).**

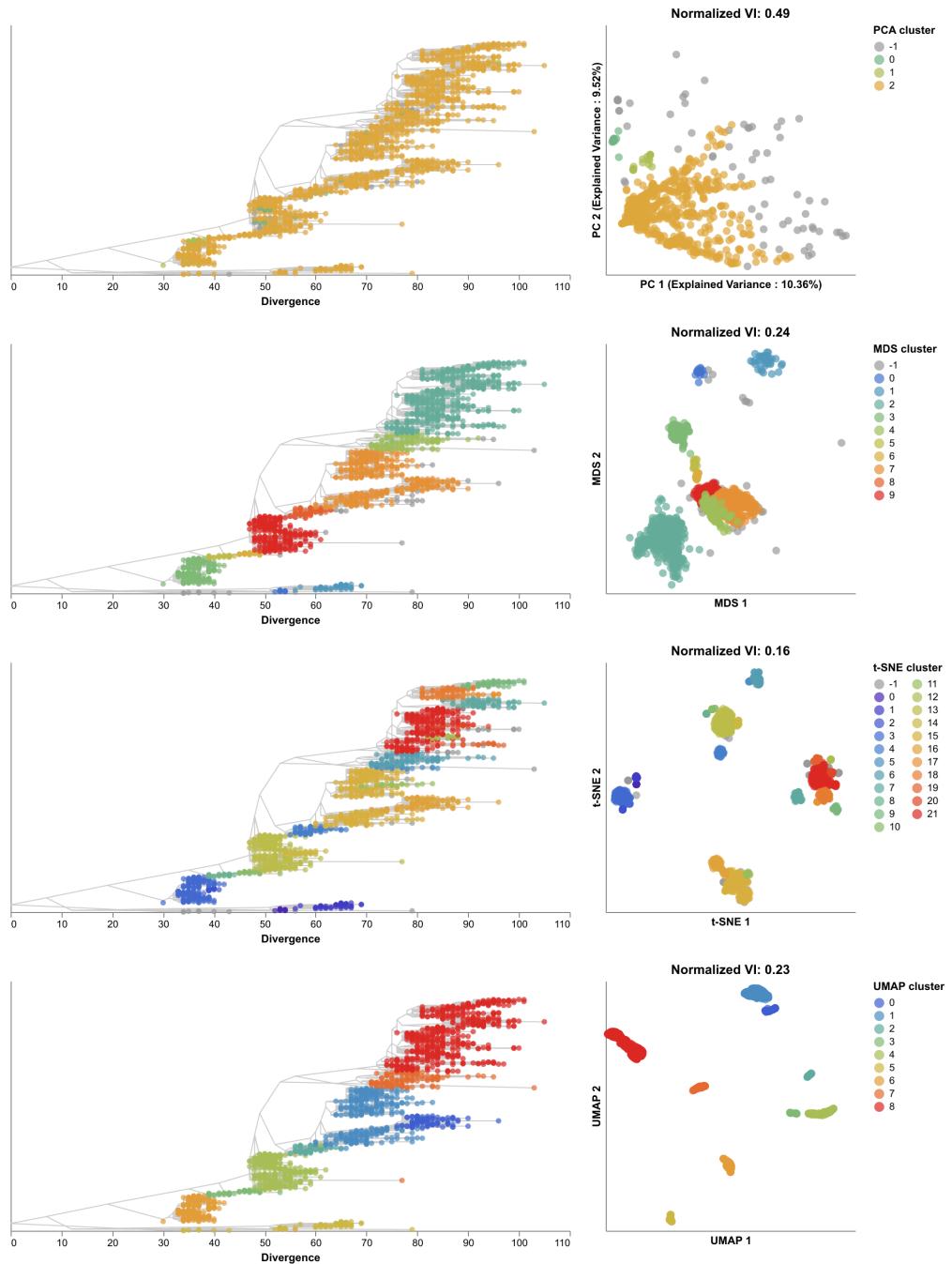
resulting embeddings could reflect pairwise genetic relationships between samples and capture previously identified genetic groups. From our analysis of simulated influenza- and coronavirus-like sequences, we found that each method produced consistent



**S18 Fig. Replication of cluster accuracy per embedding method for late (2022–2023) SARS-CoV-2 sequences across different sequences per group sampled from the original dataset and five replicates per sampling density.**

embeddings of genetic sequences for two distinct pathogens, more than 55 years of evolution, and a wide range of practical method parameters. These results suggest that researchers could apply these biologically-uninformed methods to a broad range of human pathogenic viruses with minimal tuning of the method parameters. Of the four methods, MDS most accurately reflected pairwise genetic distances between simulated samples in its embeddings. From our analysis of natural populations of seasonal influenza H3N2 HA and SARS-CoV-2 sequences, we confirmed that MDS most reliably reflected pairwise genetic distances and we found that clusters from t-SNE embeddings most accurately recapitulated previously defined genetic groups at the resolution of WHO and Nextstrain clades. Clusters from both MDS and t-SNE embeddings of H3N2 HA and NA sequences accurately matched reassortment clades identified by a biologically-informed model based on ancestral reassortment graphs. MDS embeddings consistently placed known recombinant lineages of SARS-CoV-2 between their parental lineages. From these results, we conclude that tree-free dimensionality reduction methods can provide valuable biological insights for human pathogenic viruses through easily interpretable visualizations of genetic relationships and the ability to account for genetic variation that phylogenetic methods cannot including indels, reassortment, and recombination.

Despite the promise of these simple methods to answer important public health



**S19 Fig. Phylogenetic trees (left) and embeddings (right) of late (2022–2023) SARS-CoV-2 sequences colored by HDBSCAN cluster.**  
 Normalized VI values per embedding reflect the distance between clusters and known genetic groups (collapsed Nextclade pango lineages).

questions about human pathogenic viruses, these methods and our analyses suffer from  
 739 inherent limitations. The lack of an underlying biological model is both a strength and  
 740 the clearest limitation of the dimensionality reduction methods we considered here. For  
 741

example, embeddings of SARS-CoV-2 genomes cannot capture the same fine-grained  
742 genetic resolution as Pangolin lineage annotations. Each method provides only a few  
743 parameters to tune its embeddings and these parameters have little effect on the  
744 qualitative outcome. Each method also suffers from specific issues in our analysis. PCA  
745 performs poorly with missing data and requires researchers to impute the missing values  
746 prior to analysis, as previously shown for Zika virus [30]. Neither t-SNE nor UMAP  
747 maintain a linear relationship between pairwise Euclidean and genetic distances across  
748 the observed range of genetic distances. As a result, viewers cannot know that samples  
749 mapping far apart in a t-SNE or UMAP embedding are as genetically distant as they  
750 appear. In maintaining a linear relationship between Euclidean and genetic distances,  
751 MDS sacrifices the ability to form more accurate genetic clusters for viruses with large  
752 genomes like SARS-CoV-2. Given these limitations of these methods, we do not expect  
753 them to replace biologically-informed methods that provide more meaningful  
754 parameters to tune their algorithms. Instead, we expect that researchers can use these  
755 methods for rapid visualization and clustering of their genome sequences as the first  
756 step prior to analysis with more sophisticated and computationally intensive algorithms.  
757

We note that our analysis reflects a small subset of human pathogen viruses and  
758 dimensionality reduction methods. We focused on analysis of two respiratory RNA  
759 viruses that contribute dramatically to seasonal human morbidity and mortality, but  
760 numerous alternative pathogens would also have been relevant subjects. For example,  
761 HIV represents a canonical example of a highly recombinant and bloodborne virus,  
762 while Zika, dengue, and West Nile viruses represent pathogens with multiple host  
763 species in a transmission chain. Similarly, we selected only four dimensionality  
764 reduction methods from myriad options that are commonly applied to genetic data [65].  
765 We chose these methods based on their wide use and availability in tools like  
766 scikit-learn [36] and to limit the dimensionality of our analyses.  
767

Some limitations noted above suggest future directions for this line of research. We  
768 provide optimal settings for each pathogen and embedding method in this study and  
769 open source tools to apply these methods to other pathogens. Researchers can easily  
770 integrate these tools into existing workflows for the genomic epidemiology of viruses and  
771 visualize the results with Nextstrain. Alternately, researchers may choose to apply  
772 similar existing tools developed for metagenomic analysis [66–69] to the analysis of viral  
773

populations. In the short term, researchers can apply the methods we describe here to  
774 seasonal influenza and SARS-CoV-2 genomes to identify biologically relevant clusters  
775 including reassortment events or recombinant lineages. In the long term, we expect  
776 researchers will benefit from expanding the breadth of dimensionality reduction  
777 methods applied to viruses and the breadth of viral diversity assessed by these methods.  
778

## Conclusion

We showed that simple dimensionality reduction methods operating on pairwise genetic  
780 differences can capture biologically-relevant clusters of phylogenetic clades, reassortment  
781 events, and patterns of recombining lineages for human pathogenic viruses. The  
782 conceptual and practical simplicity of these tools should enable researchers to more  
783 readily visualize and compare samples for human pathogenic viruses when phylogenetic  
784 methods are either unnecessary or inappropriate.  
785

## Supporting information

**S1 Fig. Distribution of mean absolute errors (MAE) between observed and  
predicted pairwise genetic distances per embedding method parameters for  
simulated influenza-like populations.** Each panel shows boxplots of MAEs for a  
787 specific embedding method (PCA, MDS, t-SNE, and UMAP) and a given combination  
788 of method parameters. Boxplots reflect median, upper and lower quartiles, and the  
789 range of values.  
790

**S2 Fig. Distribution of mean absolute errors (MAE) between observed and  
predicted pairwise genetic distances per embedding method parameters for  
simulated coronavirus-like populations.** Each panel shows boxplots of MAEs for a  
793 specific embedding method (PCA, MDS, t-SNE, and UMAP) and a given combination  
794 of method parameters. Boxplots reflect median, upper and lower quartiles, and the  
795 range of values.  
796

**S3 Fig. Representative MDS embeddings for simulated populations using  
optimal parameters per pathogen (rows) and showing all three components.**  
797  
798

S4 Fig. Pairwise nucleotide distances for early (2016–2018) and late  
801  
(2018–2020) influenza H3N2 HA sequences within and between genetic  
802 groups defined by Nextstrain clades and clusters from PCA, MDS, t-SNE,  
803 and UMAP embeddings.  
804

S5 Fig. Phylogeny of early (2018–2020) influenza H3N2 HA sequences  
805 plotted by nucleotide substitutions per site on the x-axis (top) and  
806 low-dimensional embeddings of the same sequences by PCA (middle left),  
807 MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips  
808 in the tree and embeddings are colored by their Nextstrain clade assignment.  
809

S6 Fig. MDS embeddings for late (2018–2020) influenza H3N2 HA  
810 sequences showing all three components.  
811

S7 Fig. Replication of cluster accuracy per embedding method for late  
812 (2018–2020) influenza H3N2 HA sequences across different sequences per  
813 group sampled from the original dataset and five replicates per sampling  
814 density.  
815

S8 Fig. Embeddings influenza H3N2 HA-only (left) and combined HA/NA  
816 (right) showing the effects of additional NA genetic information on the  
817 placement of reassortment events detected by TreeKnit (MCCs).  
818

S9 Fig. PCA embeddings for influenza H3N2 HA sequences only (top row)  
819 and HA/NA sequences combined (bottom row) showing the HA trees  
820 colored by clusters identified in each embedding (left) and the  
821 corresponding embeddings colored by cluster (right).  
822

S10 Fig. MDS embeddings for influenza H3N2 HA sequences only (top  
823 row) and HA/NA sequences combined (bottom row) showing the HA trees  
824 colored by clusters identified in each embedding (left) and the  
825 corresponding embeddings colored by cluster (right).  
826

- S11 Fig. t-SNE embeddings for influenza H3N2 HA sequences only (top 827  
row) and HA/NA sequences combined (bottom row) showing the HA trees 828  
colored by clusters identified in each embedding (left) and the 829  
corresponding embeddings colored by cluster (right). 830
- S12 Fig. UMAP embeddings for influenza H3N2 HA sequences only (top 831  
row) and HA/NA sequences combined (bottom row) showing the HA trees 832  
colored by clusters identified in each embedding (left) and the 833  
corresponding embeddings colored by cluster (right). 834
- S13 Fig. MDS embeddings for early SARS-CoV-2 sequences showing all 835  
three components. 836
- S14 Fig. Principal component 1 (PC1) of the PCA embedding for early 837  
SARS-CoV-2 data plotted by the number of missing (“N”) or gap (“-”) 838  
characters in the corresponding sample’s aligned sequence. Pearson’s  $R^2$  839  
estimates the variation in PC1 explained by missing data. 840
- S15 Fig. Phylogeny of early (2020–2022) SARS-CoV-2 sequences plotted 841  
by number of nucleotide substitutions from the most recent common 842  
ancestor on the x-axis (top) and low-dimensional embeddings of the same 843  
sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), 844  
and UMAP (bottom right). Tips in the tree and embeddings are colored by 845  
their collapsed Nextclade pango lineage assignment. 846
- S16 Fig. Pairwise nucleotide distances for early (2020–2022) and late 847  
(2022–2023) SARS-CoV-2 sequences within and between genetic groups 848  
defined by Nextstrain clades, collapsed Nextclade pango lineages, and 849  
clusters from PCA, MDS, t-SNE, and UMAP embeddings. 850
- S17 Fig. Phylogenetic trees (left) and embeddings (right) of early 851  
(2020–2022) SARS-CoV-2 sequences colored by HDBSCAN cluster. 852  
Normalized VI values per embedding reflect the distance between clusters 853  
and known genetic groups (collapsed Nextclade pango lineages). 854

S18 Fig. Replication of cluster accuracy per embedding method for late (2022–2023) SARS-CoV-2 sequences across different sequences per group sampled from the original dataset and five replicates per sampling density.	855
S19 Fig. Phylogenetic trees (left) and embeddings (right) of late (2022–2023) SARS-CoV-2 sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (collapsed Nextclade pango lineages).	858 859 860 861
S1 Table. Mutations observed per embedding cluster relative to a reference genome sequence for each pathogen. Each row reflects the alternate allele identified at a specific position of the given pathogen genome or gene sequence, the pathogen dataset, the embedding method, the number of clusters in the embedding with the observed mutation, and the list of distinct cluster labels with the mutation. Mutations must have occurred in at least 10 samples of the given dataset with an allele frequency of at least 50%.	862 863 864 865 866 867 868 869
S2 Table. Average Euclidean distances between each known recombinant, “X”, and its parental lineages “A” and “B” per embedding method. Distances include average pairwise comparisons between A and B, A and X, and B and X. Additional columns indicate whether each recombinant lineage maps closer to both parental lineages (or at least one) than those parents map to each other.	870 871 872 873 874 875
S3 Table. Accessions and authors from originating and submitting laboratories of seasonal influenza and SARS-CoV-2 sequences from INSDC databases.	876 877 878

## Acknowledgments

We thank members of the Bedford Lab for constructive feedback on this project over the course of many years. We gratefully acknowledge the originating and submitting

## References

1. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, et al. Tracking virus outbreaks in the twenty-first century. *Nat Microbiol.* 2019;4(1):10–19.
2. Volz EM, Koelle K, Bedford T. Viral phylodynamics. *PLoS Comput Biol.* 2013;9(3):e1002947.
3. Baele G, Suchard MA, Rambaut A, Lemey P. Emerging Concepts of Data Integration in Pathogen Phylodynamics. *Syst Biol.* 2017;66(1):e47–e65.
4. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution.* 2018;4(1). doi:10.1093/ve/vex042.
5. Nelson MI, Viboud C, Simonsen L, Bennett RT, Griesemer SB, St George K, et al. Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog.* 2008;4(2):e1000012.
6. Marshall N, Priyamvada L, Ende Z, Steel J, Lowen AC. Influenza virus reassortment occurs with high frequency in the absence of segment mismatch. *PLoS Pathog.* 2013;9(6):e1003421.
7. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol.* 2016;24(6):490–502.
8. Lemey P, Kosakovsky Pond SL, Drummond AJ, Pybus OG, Shapiro B, Barroso H, et al. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput Biol.* 2007;3(2):e29.
9. Bhatt S, Holmes EC, Pybus OG. The genomic rate of molecular adaptation of the human influenza A virus. *Mol Biol Evol.* 2011;28(9):2443–2451.

10. Wiens JJ. Combining data sets with different phylogenetic histories. *Syst Biol.* 1998;47(4):568–581.
11. Barrat-Charlaix P, Vaughan TG, Neher RA. TreeKnit: Inferring ancestral reassortment graphs of influenza viruses. *PLoS Computational Biology.* 2022;18(8):e1010394.
12. Muller NF, Kistler KE, Bedford T. A Bayesian approach to infer recombination patterns in coronaviruses. *Nat Commun.* 2022;13(1):4186.
13. O'Toole A, Hill V, Jackson B, Dewar R, Sahadeo N, Colquhoun R, et al. Genomics-informed outbreak investigations of SARS-CoV-2 using civet. *PLOS Glob Public Health.* 2022;2(12):e0000704.
14. McBroome J, Martin J, de Bernardi Schneider A, Turakhia Y, Corbett-Detig R. Identifying SARS-CoV-2 regional introductions and transmission clusters in real time. *Virus Evol.* 2022;8(1):veac048.
15. Stoddard G, Black A, Ayscue P, Lu D, Kamm J, Bhatt K, et al. Using genomic epidemiology of SARS-CoV-2 to support contact tracing and public health surveillance in rural Humboldt County, California. *BMC Public Health.* 2022;22(1):456.
16. Tran-Kiem C, Bedford T. Estimating the reproduction number and transmission heterogeneity from the size distribution of clusters of identical pathogen sequences. *medRxiv.* 2023;doi:10.1101/2023.04.05.23287263.
17. O'Toole A, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* 2021;7(2):veab064.
18. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat Genet.* 2021;53(6):809–816.
19. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software.* 2021;6(67):3773. doi:10.21105/joss.03773.

20. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom*. 2016;2(11):e000093.
21. Campbell EM, Boyles A, Shankar A, Kim J, Knyazev S, Cintron R, et al. MicrobeTrace: Retooling molecular epidemiology for rapid public health response. *PLoS Comput Biol*. 2021;17(9):e1009300.
22. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philosophical transactions Series A, Mathematical, physical, and engineering sciences*. 2016;;
23. Hout MC, Papesh MH, Goldinger SD. Multidimensional scaling. Wiley Online Library. 2012;;
24. Maaten Lvd, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research*. 2008;9(Nov):2579–2605.
25. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018;.
26. McVean G. A genealogical interpretation of principal components analysis. *PLoS Genet*. 2009;5(10):e1000686.
27. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. *Nature*. 2008;.
28. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*. 2009;.
29. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.
30. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, et al. Zika virus evolution and spread in the Americas. *Nature*. 2017;546(7658):411–415.
31. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. *Nature*. 2008;.

32. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat Biotechnol.* 2021;39(2):156–157.
33. Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genetics.* 2019;;
34. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol.* 2018;;
35. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun.* 2019;10(1):5416.
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011;12:2825–2830.
37. Jacobs RA. Increased rates of convergence through learning rate adaptation. *Neural Networks.* 1988;1(4):295–307.  
doi:[https://doi.org/10.1016/0893-6080\(88\)90003-2](https://doi.org/10.1016/0893-6080(88)90003-2).
38. Belkina AC, Ciccolella CO, Anno R, Halpert R, Spidlen J, Snyder-Cappione JE. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat Commun.* 2019;10(1):5415.
39. Jariani A, Warth C, Deforche K, Libin P, Drummond AJ, Rambaut A, et al. SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination. *Virus Evolution.* 2019;5(1).
40. Huddleston J, Barnes JR, Rowe T, Xu X, Kondor R, Wentworth DE, et al. Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. *eLife.* 2020;9:e60067. doi:10.7554/eLife.60067.
41. Rambaut A. Phylogenetic analysis of nCoV-2019 genomes. *Virological;*

42. Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, de Silva TI, et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology*. 2023;doi:10.1038/s41579-022-00841-7.
43. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 3rd ed. Melbourne, Australia: OTexts; 2021. Available from: [OTexts.com/fpp3](http://OTexts.com/fpp3).
44. Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. *Nucleic Acids Res*. 2021;49(D1):D121–D124.
45. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*. 2002;30(14):3059–3066. doi:10.1093/nar/gkf436.
46. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772–780.
47. Huddleston J, Hadfield J, Sibley TR, Lee J, Fay K, Ilcisin M, et al. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. *J Open Source Softw*. 2021;6(57).
48. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*. 2014;32(1):268–274. doi:10.1093/molbev/msu300.
49. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018; p. bty407. doi:10.1093/bioinformatics/bty407.
50. Neher RA, Bedford T. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*. 2015;31(21):3546–3548.
51. Hodcroft EB, J H, A NR, Bedford T. Year-letter Genetic Clade Naming for SARS-CoV-2 on Nextstrain.org; 2020.  
<https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming>.

52. Bedford T, Hodcroft EB, A NR. Updated Nextstrain SARS-CoV-2 clade naming strategy; 2021. <https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming>.
53. Roemer C, Hodcroft EB, A NR, Bedford T. SARS-CoV-2 clade naming strategy for 2022; 2022. <https://nextstrain.org/blog/2022-04-29-SARS-CoV-2-clade-naming-2022>.
54. Campello RJ, Moulavi D, Zimek A, Sander J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2015;10(1):1–51.
55. Meilă M. Comparing clusterings by the variation of information. In: Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings. Springer; 2003. p. 173–187.
56. Mölder F, Jablonski K, Letcher B, Hall M, Tomkins-Tinch C, Sochat V, et al. Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]. *F1000Research*. 2021;10(33). doi:10.12688/f1000research.29032.2.
57. Petrova VN, Russell CA. The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*. 2018;16(1):47–60. doi:10.1038/nrmicro.2017.118.
58. Hay AJ, McCauley JW. The WHO global influenza surveillance and response system (GISRS)-A future perspective. *Influenza Other Respir Viruses*. 2018;12(5):551–557.
59. Potter BI, Kondor R, Hadfield J, Huddleston J, Barnes J, Rowe T, et al. Evolution and rapid spread of a reassortant A(H3N2) virus that predominated the 2017–2018 influenza season. *Virus Evolution*. 2019;5(2). doi:10.1093/ve/vez046.
60. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*. 2020;382(8):727–733. doi:10.1056/NEJMoa2001017.

61. Kistler KE, Huddleston J, Bedford T. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. *Cell Host Microbe*. 2022;30(4):545–555.
62. Focosi D, Maggi F. Recombination in Coronaviruses, with a Focus on SARS-CoV-2. *Viruses*. 2022;14(6).
63. Turakhia Y, Thornlow B, Hinrichs A, McBroome J, Ayala N, Ye C, et al. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature*. 2022;609(7929):994–997.
64. Konings F, Perkins MD, Kuhn JH, Pallen MJ, Alm EJ, Archer BN, et al. SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nat Microbiol*. 2021;6(7):821–823.
65. Armstrong G, Rahman G, Martino C, McDonald D, Gonzalez A, Mishne G, et al. Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data. *Front Bioinform*. 2022;2:821861.
66. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75(23):7537–7541.
67. Schloss PD. Reintroducing mothur: 10 Years Later. *Appl Environ Microbiol*. 2020;86(2).
68. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*. 2019;37(8):852–857.
69. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*. 2013;8(4):e61217.