Dimensionality reduction distills complex evolutionary relationships in seasonal influenza and SARS-CoV-2

Sravani Nanduri[1], Allison Black[2], Trevor Bedford[2,3], John Huddleston[2*]

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 Center, Seattle, WA, USA

3 Howard Hughes Medical Institute, Seattle, WA, USA

* jhuddles@fredhutch.org

Abstract

Public health researchers and practitioners 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 or difficult to interpret without specialty knowledge. 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 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 and best captured the intermediate placement of recombinant lineages between parental lineages. Clusters from t-SNE most accurately recapitulated known phylogenetic clades and recombinant lineages. Both MDS and t-SNE accurately identified reassortment groups. 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

To track the progress of viral epidemics, public health researchers often need to identify groups of genetically-related samples. A common approach to find these groups involves inferring the complete evolutionary history of virus samples using phylogenetic methods. However, these methods assume that new viruses descend from a single parent, while many viruses including seasonal influenza and SARS-CoV-2 produce offspring through a form of sexual reproduction that violates this assumption. Additionally, phylogenies may be unnecessarily complex or unintuitive when researchers only need to find and visualize clusters of related samples. We tested an alternative approach by applying widely-used statistical methods (PCA, MDS, t-SNE, and UMAP) to create 2- or 3-dimensional maps of virus samples from their pairwise genetic distances and identify clusters of samples that place close together in these maps. We found that these statistical methods without an underlying biological model could accurately capture known genetic relationships in populations of seasonal influenza and SARS-CoV-2 even in the presence of sexual reproduction. The conceptual and practical simplicity of our open source implementation of these methods enables researchers to visualize and compare human pathogenic virus samples when phylogenetic methods are unnecessary or inappropriate.

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 the sequence diversity of sampled tips accrued through clonal evolution, that is, the occurrence of mutations on top of an inherited genomic background, that is further inherited by descendent pathogens. 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 recombination graphs [11, 12]. However, several key questions in genomic epidemiology do not require inference of ancestral relationships and states, and therefore may be amenable to non-phylogenetic approaches for summarizing genetic relationships. For example, genomic epidemiologists commonly need to 1) visualize the genetic relationships among closely related virus samples [13, 14], 2) identify clusters of closely-related genomes that represent regional outbreaks or new variants of concern [15–18], 3) place newly sequenced viral genomes in the evolutionary context of other circulating samples [19–21]. Given that these common use cases rely on genetic distances between samples, tree-free statistical methods that operate on pairwise distances could be sufficient to address each case. As these tree-free methods lack a formal biological model of evolutionary relationships, they make weak assumptions about the input data and therefore should be applicable to pathogen genomes that violate phylogenetic assumptions. Furthermore, methods that describe genetic relationships with map-like visualizations may feel more familiar to public health practitioners, and therefore more easily applied for public health action.

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

Each of the embedding methods mentioned above has been applied previously to genomic data to visualize relationships between individuals and identify clusters of related genomes. Although PCA is a generic linear algebra algorithm that optimizes for an orthogonal embedding of the data, the principal components from single nucleotide polymorphisms (SNPs) represent mean coalescent times and therefore recapitulate broad phylogenetic relationships [26]. PCA has been applied to SNPs of human genomes [26–29] and to multiple sequence alignments of viral genomes [30]. MDS attempts to embed input data into a lower-dimensional representation such that each pair of data points are as close in the embedding as they are in the original high-dimensional space. MDS has been applied to multiple gene segments of seasonal influenza viruses to visualize evolutionary relationships between segments [31] and to individual influenza gene segments to reveal low-dimensional trajectories of genetic clusters [32, 33]. Both t-SNE and UMAP build on manifold learning methods like MDS to find low-dimensional embeddings of data that place similar points close together and

dissimilar points far apart [34]. These methods have been applied to SNPs from human genomes [35] and single-cell transcriptomes [36, 37].

Although these methods are commonly used for qualitative studies of evolutionary relationships, few studies have attempted to quantify patterns observed in the resulting embeddings, investigate the value of applying these methods to viruses that reassort or recombine, or identify optimal method parameters for application to viruses. Recent studies disagree about whether methods like PCA, t-SNE, and UMAP produce meaningful global structures [34] or arbitrary patterns that distort high-dimensional relationships [38]. To address these open questions, we tuned and validated the performance of PCA, MDS, t-SNE, and UMAP with genomes from simulated influenza-like and coronavirus-like populations and then applied these methods to natural populations of seasonal influenza virus A/H3N2 and SARS-CoV-2. These natural viruses are highly relevant as major causes of global human mortality, common subjects of real-time genomic epidemiology, and representatives of reassortant and recombinant human pathogens. For each combination of virus and embedding method, we quantified the relationship between pairwise genetic and Euclidean embedding distances, identified clusters of closely-related genomes in embedding space, and evaluated the accuracy of clusters compared to genetic groups defined by experts. Finally, we tested the ability of these methods to capture patterns of reassortment between seasonal influenza A/H3N2 hemagglutinin (HA) and neuraminidase (NA) segments and recombination in SARS-CoV-2 genomes. These results inform our recommendations for future applications of these methods including which are most effective for specific problems in genomic epidemiology and which parameters researchers should use for each method.

Results

The ability of embedding methods to produce global structures for simulated viral populations varies little across method parameters

To understand how well PCA, MDS, t-SNE, and UMAP could represent genetic relationships between samples of human pathogenic viruses under well-defined evolutionary conditions, we simulated influenza-like and coronavirus-like populations and created embeddings for each population across a range of method parameters. We maximized the local and global interpretability of each method's embeddings by identifying parameters that maximized a linear relationship between genetic distance and Euclidean distance in low-dimensional space (see Methods). Specifically, we selected parameters that minimized the median of the mean absolute error (MAE) between observed pairwise genetic distances of simulated genomes and predicted genetic distances for those genomes based on their Euclidean distances in each embedding. For methods like PCA and MDS where increasing the number of components available to the embedding could lead to overfitting, we selected the maximum number of components beyond which the median MAE did not decrease by more than 1 nucleotide.

For influenza-like populations, the optimal parameters were 2 components for PCA, 3 components for MDS, perplexity of 100 and learning rate of 100 for t-SNE, and nearest neighbors of 100 and minimum distance of 0.1 for UMAP. As expected, increasing the number of components for PCA and MDS gradually decreased the median MAEs of their embeddings (S1 Fig A and B). However, beyond 2 and 3 components, respectively, the reduction in error did not exceed 1 nucleotide. This result suggests that there were diminishing returns for the increased complexity of additional components. Both t-SNE and UMAP embeddings produced a wide range of errors (the majority between 10 and 20 average mismatches) across all parameter values (S1 Fig C and D). Embeddings from t-SNE appeared robust to variation in parameters, with a slight improvement in median MAE associated with perplexity of 100 and little benefit to any of the learning rate values (S1 Fig C). Similarly, UMAP embeddings were robust across the range of tested parameters, with the greatest benefit coming from setting the nearest neighbors greater

than 25 and no benefit from changing the minimum distance between points (S1 Fig D).

The optimal parameters for coronavirus-like populations were similar to those for the influenza-like populations. The optimal parameters were 2 components for PCA, 3 for MDS, perplexity of 100 and learning rate of 500 for t-SNE, and nearest neighbors of 50 and minimum distance of 0.1 for UMAP. As with influenza-like populations, both PCA and MDS showed diminishing benefits of increasing the number of components (S2 Fig A and B). Similarly, we observed little improvement in MAEs from varying t-SNE and UMAP parameters (S2 Fig C and D). The most noticeable improvement came from setting t-SNE's perplexity to 100 (S2 Fig C). These results indicate the limits of t-SNE and UMAP to represent global genetic structure from these data.

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

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 [39]. These mutations produce distinct phylogenetic clades that represent potentially different antigenic phenotypes. The World Health Organization (WHO) Global Influenza Surveillance and Response System regularly sequences genomes of circulating influenza lineages [40] and submits these

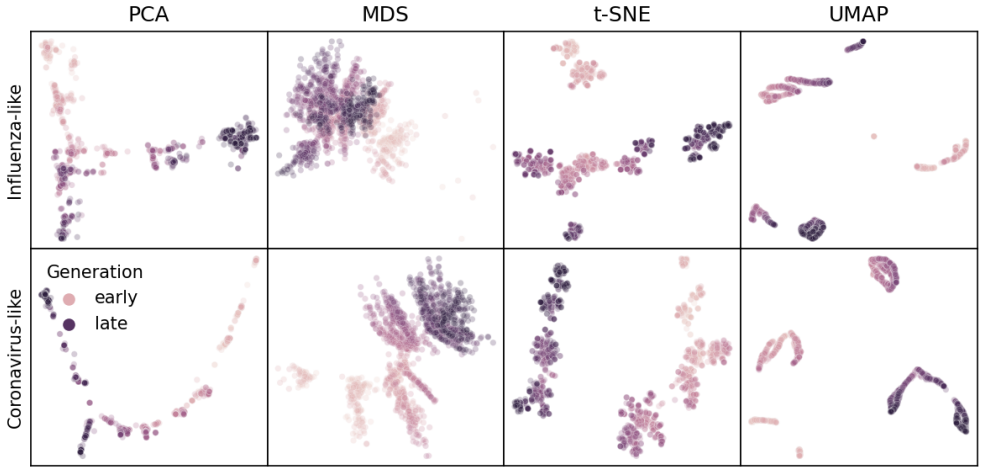

Fig 1. Representative embeddings for simulated populations using optimal parameters per pathogen (rows) and embedding method (columns). Each panel shows the embedding for sequences from the first four years of a single replicate population for the corresponding pathogen type. Each point represents a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

sequences to public INSDC databases like NCBI's GenBank [41]. These factors, coupled with HA's relatively short gene size of 1,701 nucleotides, facilitate real-time genomic epidemiology of H3N2 [42] and rapid analysis by the embedding methods we wanted to evaluate. We analyzed H3N2 HA sequences from two consecutive time periods including an "early" dataset from 2016–2018 and a "late" dataset from 2018–2020. For each dataset, we created embeddings with all four methods, identified clusters in the embeddings with HDBSCAN, and calculated the accuracy of clusters relative to expert-defined genetic groups (see Methods). We used the early dataset to identify cluster parameters that minimized the distance between clusters and known genetic groups. We tested these optimal parameters with the late dataset. This approach allowed us to maximize cluster accuracy against the background of embedding method parameters that we already optimized to maximize interpretability of visualizations.

We first applied each embedding method to the early H3N2 HA sequences (2016–2018) and compared the placement of these sequences in the embeddings to their corresponding clades in the phylogeny. All four embedding methods qualitatively recapitulated the 10 Nextstrain clades observed in the phylogeny (Fig 2 and S4 Fig). Samples from the same clade generally grouped tightly together. Most embedding

methods also delineated larger phylogenetic clades, placing clades A1, A2, A3, A4, and 3c3.A into separate locations in the embeddings. Despite maintaining local and broader global structure, not all embeddings captured intermediate genetic structure. For example, all methods placed A1b and its descendant clades, A1b/135K and A1b/135N, into tight clusters together. The t-SNE embedding created separate clusters for each of these 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 methods maintained a linear pairwise relationship for samples that differed by no more than ≈10 nucleotides (Fig 3). Only MDS consistently maintained that linearity as genetic distance increased (Pearson's $R^2 = 0.94$). We observed a less linear relationship for samples with more genetic differences in PCA (Pearson's $R^2 = 0.67$), t-SNE (Pearson's $R^2 = 0.34$), and UMAP (Pearson's $R^2 = 0.68$) embeddings. While PCA and UMAP Euclidean distances increased monotonically with genetic distance, t-SNE embeddings placed some pairs of samples with intermediate distances of 30-40 nucleotides farther apart than pairs of samples with much greater genetic distances.

Next, we found clusters in embeddings of early H3N2 HA data and calculated their distance to previously defined genetic groups. We assigned cluster labels to each sample with the hierarchical clustering algorithm, HDBSCAN [43]. We calculated distances between clusters and known genetic groups with the normalized variation of information (VI) metric [44] which produces a value of 0 for identical groups and 1 for maximally different groups (see Methods). HDBSCAN does not require an expected number of clusters as input, but it does provide a parameter for the minimum distance required between clusters. We optimized this minimum distance threshold by minimizing the VI distance between known genetic groups and clusters produced with different threshold values (S1 Table). Clusters produced with the optimal distance threshold were generally monophyletic (S2 Table), supported by cluster-specific mutations (S3 Table), and corresponded to larger phylogenetic clades (Fig 4). Pairwise genetic distances between

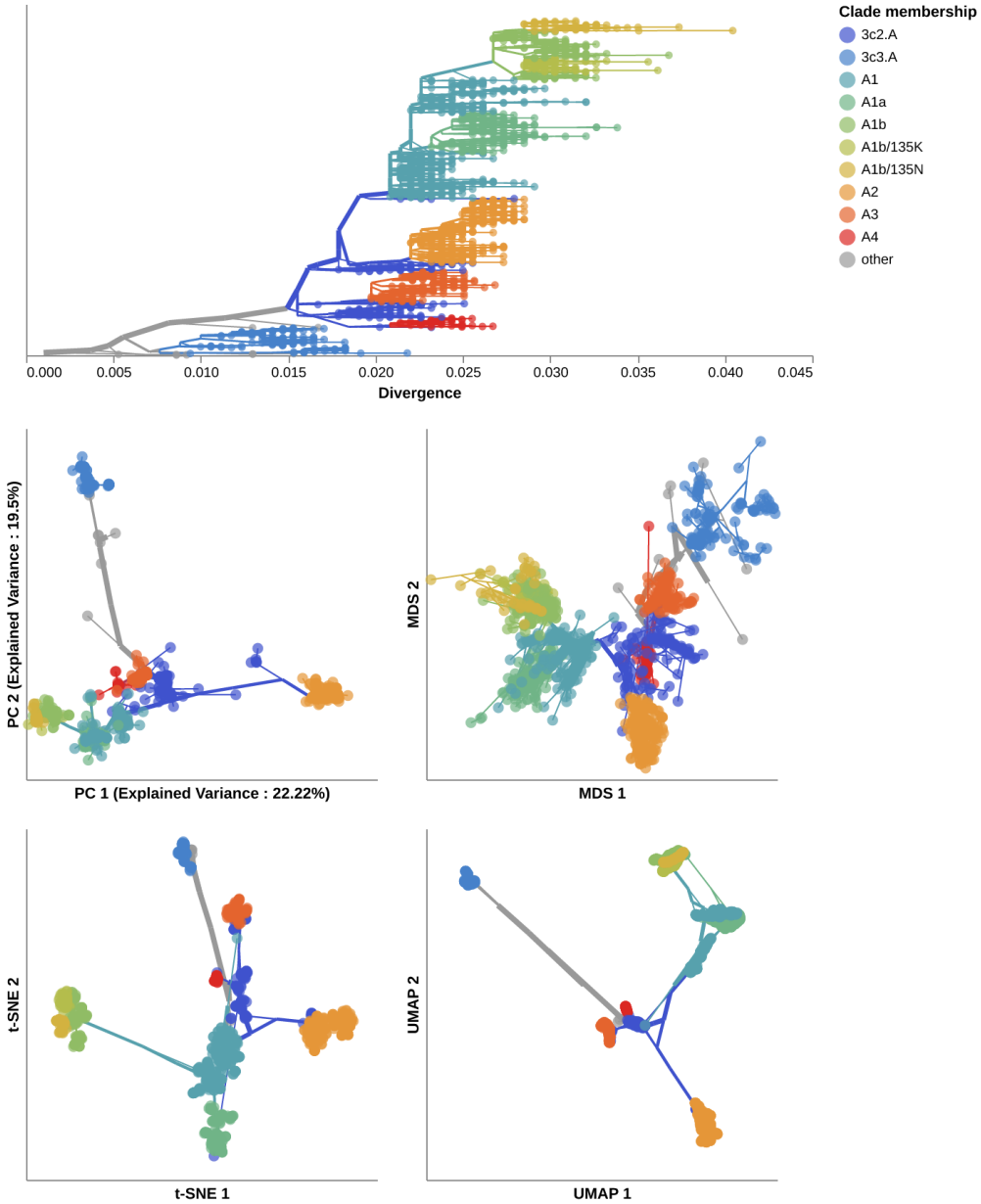

Fig 2. Phylogeny of early (2016–2018) influenza H3N2 HA sequences plotted by nucleotide substitutions per site on the x-axis (top) and low-dimensional embeddings of the same sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips in the tree and embeddings are colored by their Nextstrain clade assignment. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line colors represent the clade membership of the most ancestral node in the pair of nodes connected by the segment. Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

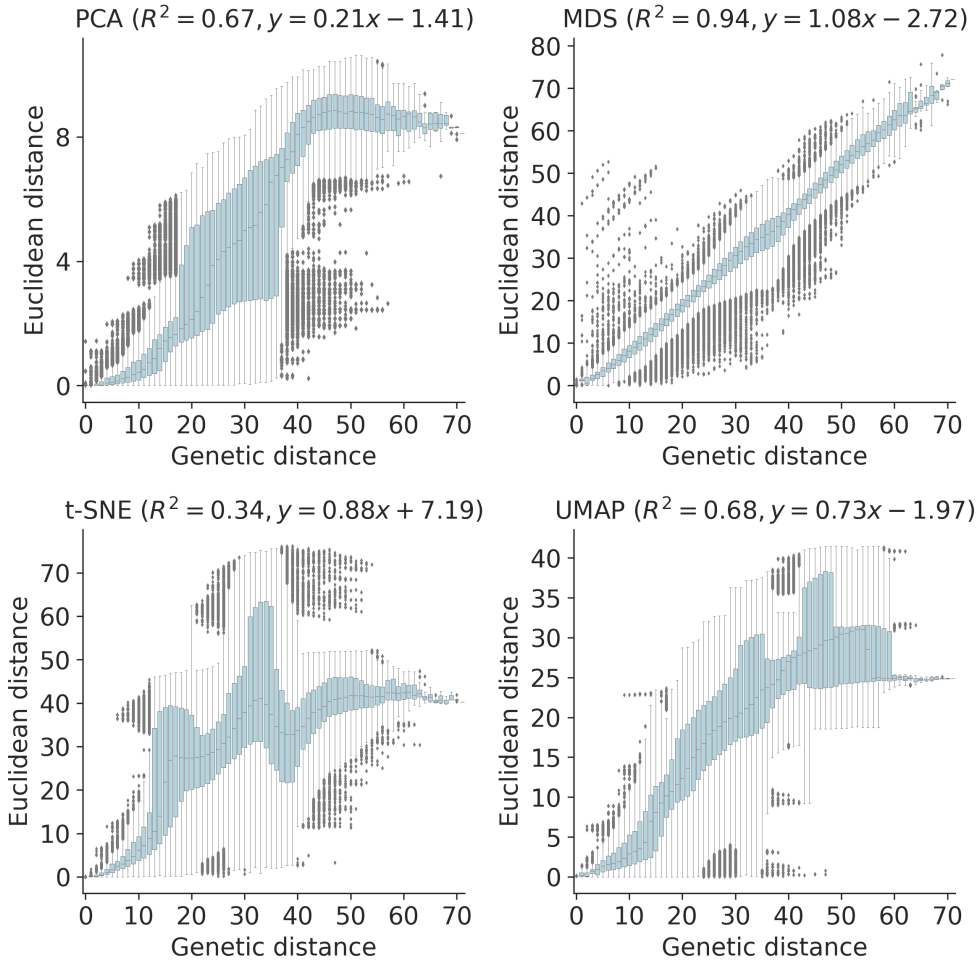

Fig 3. Relationship between pairwise genetic and Euclidean distances in embeddings of early (2016–2018) influenza H3N2 HA sequences by PCA (upper left), MDS (upper right), t-SNE (lower left), and UMAP (lower right). Each boxplot represents the distribution of pairwise Euclidean distances at a given genetic distance. Panel titles include Pearson's R^2 values and linear regression coefficients between the plotted distances.

sequences in the same MDS, t-SNE, or UMAP clusters matched the distances between sequences within Nextstrain clades (S5 Fig). The 8 clusters from t-SNE most accurately captured expert clade assignments (normalized VI=0.04) followed by UMAP's 7 clusters (normalized VI=0.09), MDS's 9 clusters (normalized VI=0.11), and PCA's 3 clusters (normalized VI=0.19). Clusters from t-SNE, MDS, and UMAP captured broader phylogenetic clades (A1, A1b, A2, A3, A4, 3c2.A, and 3c3.A) but failed to distinguish between A1b and its descendants. PCA clusters corresponded to the most distantly-related and ancestral clades (3c2.A, 3c3.A, and A2). These results indicate

that nonlinear t-SNE embeddings could be better-suited for clustering and classification than the more linear embeddings from PCA, MDS, and UMAP.

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), identified clusters per embedding, and calculated the VI distance between clusters and previously defined clades. The late dataset included 9 clades with at least 10 samples (S6 Fig). These clades had a greater average between-clade distance than clades in the early dataset (S5 Fig). As with the early dataset, clusters from the late dataset were largely monophyletic (S2 Table), supported by cluster-specific mutations (S3 Table), and corresponded to larger phylogenetic clades (Fig. 5 and S6 Fig) Pairwise genetic distances within clusters generally matched the diversity within Nextstrain clades (S5 Fig). Clusters from PCA (N=6), MDS (N=6), t-SNE (N=5), and UMAP (N=8) were similarly accurate, with normalized VI distances of 0.09, 0.07, 0.08, and 0.06, respectively (Fig. 5 and S7 Fig). MDS split A3 samples into two widely separated groups in its Euclidean space, indicating substantial within-clade genetic differences. We found recurrent HA1 substitutions of 135K, 142G, and 193S in multiple subclades of A3 that MDS could not effectively represent. Cluster accuracies were robust to changes in sampling density under the same even geographic and temporal sampling scheme, with PCA and MDS clusters producing the lowest median distance to Nextstrain clades (S8 Fig A). However, biased sampling toward the USA and clade 3c3.A decreased cluster accuracy for t-SNE and UMAP (S8 Fig B). These results show that all four methods can produce clusters that accurately capture known genetic groups when applied to previously unseen H3N2 HA samples with unbiased sampling.

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 [39] and HA and NA genes frequently reassort [5, 6, 45].

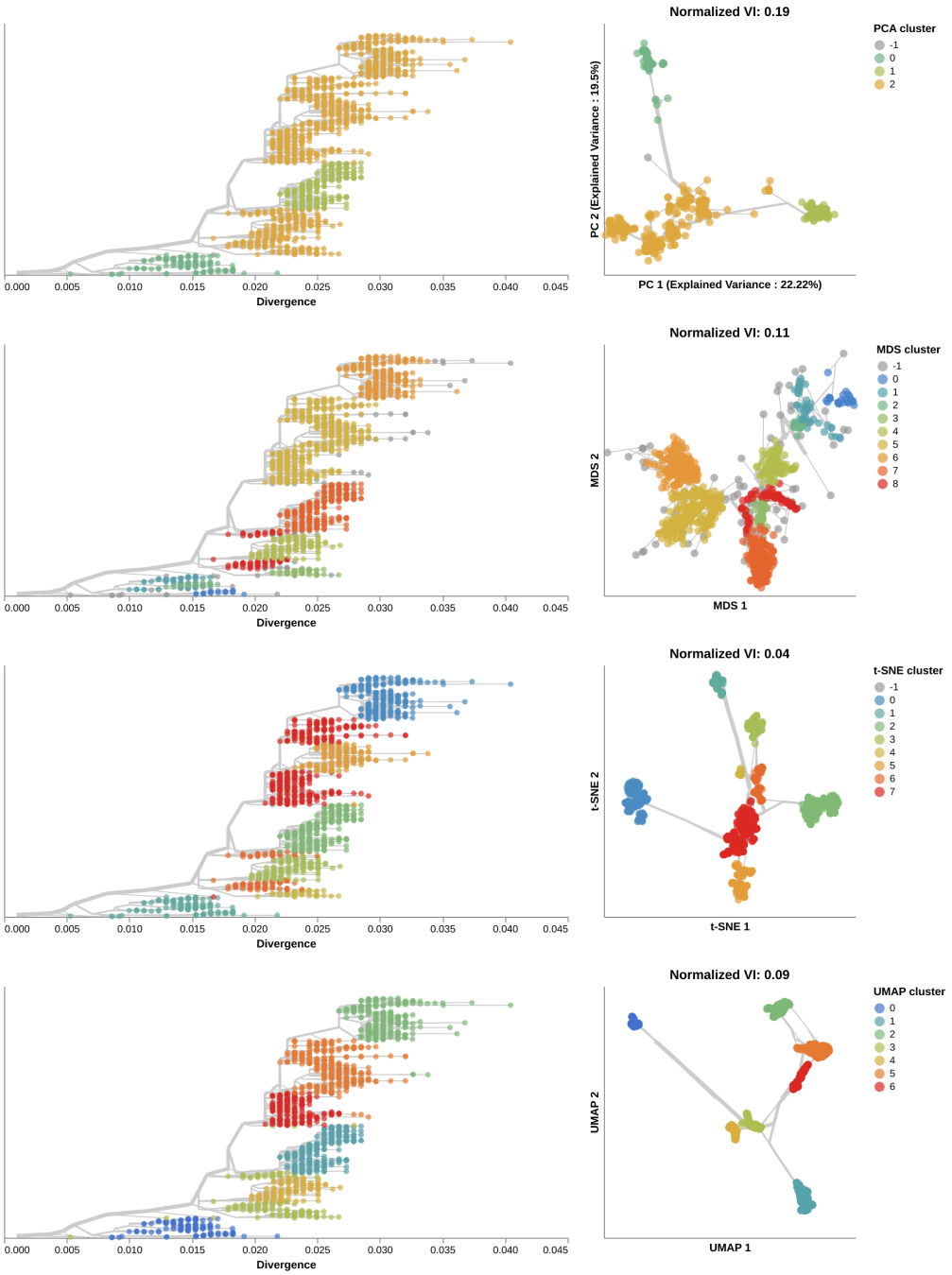

Fig 4. Phylogenetic trees (left) and embeddings (right) of early (2016–2018) influenza H3N2 HA sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Nextstrain clades). Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

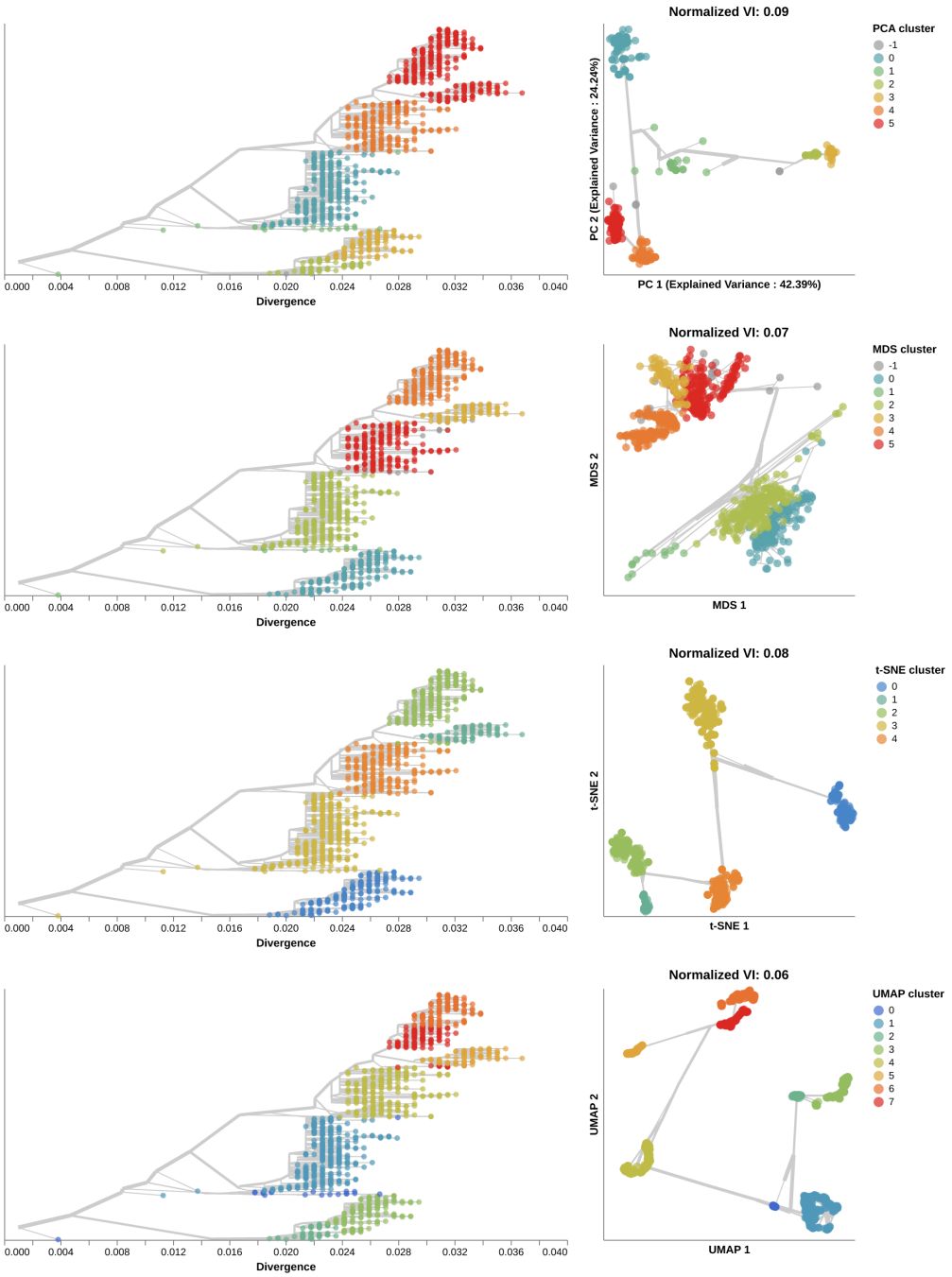

Fig 5. Phylogenetic trees (left) and embeddings (right) of late (2018–2020) H3N2 HA sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Nextstrain clades). Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

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 inferred HA and NA phylogenies from these sequences and applied TreeKnit to both trees to identify maximally compatible clades (MCCs) that represent reassortment events [11]. Of the 208 reassortment events identified by TreeKnit, 15 (7%) contained at least 10 samples representing 1,049 samples (65%).

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 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.

All embedding methods produced more accurate clusters from the HA/NA alignments than the HA-only alignments (Fig. 6 and S9 Fig). HA/NA clusters from MDS reduced the distance to known reassortment events from a normalized VI value of 0.17 with HA only to 0.06. Similarly, HA/NA clusters from t-SNE reduced the distance from 0.11 to 0.06. Adding NA to HA only modestly improved PCA and UMAP clusters, reducing distances by 0.05 and 0.03, respectively. Embeddings with both genes produced more clusters in PCA, MDS, and t-SNE than the HA-only embeddings with 1 additional cluster in PCA (S10 Fig), 9 in MDS (S11 Fig), 6 in t-SNE (S12 Fig), and 0 in UMAP (S13 Fig). With the exception of PCA, all embeddings of HA/NA alignments produced distinct clusters for the known reassortment event within clade A2 [45] as represented by MCCs 14 and 11. Other larger events like those represented by MCCs 9 and 12 mapped far apart in all HA/NA embeddings except PCA. We noted that some of the additional clusters in HA/NA embeddings likely also reflected genetic diversity in NA 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.

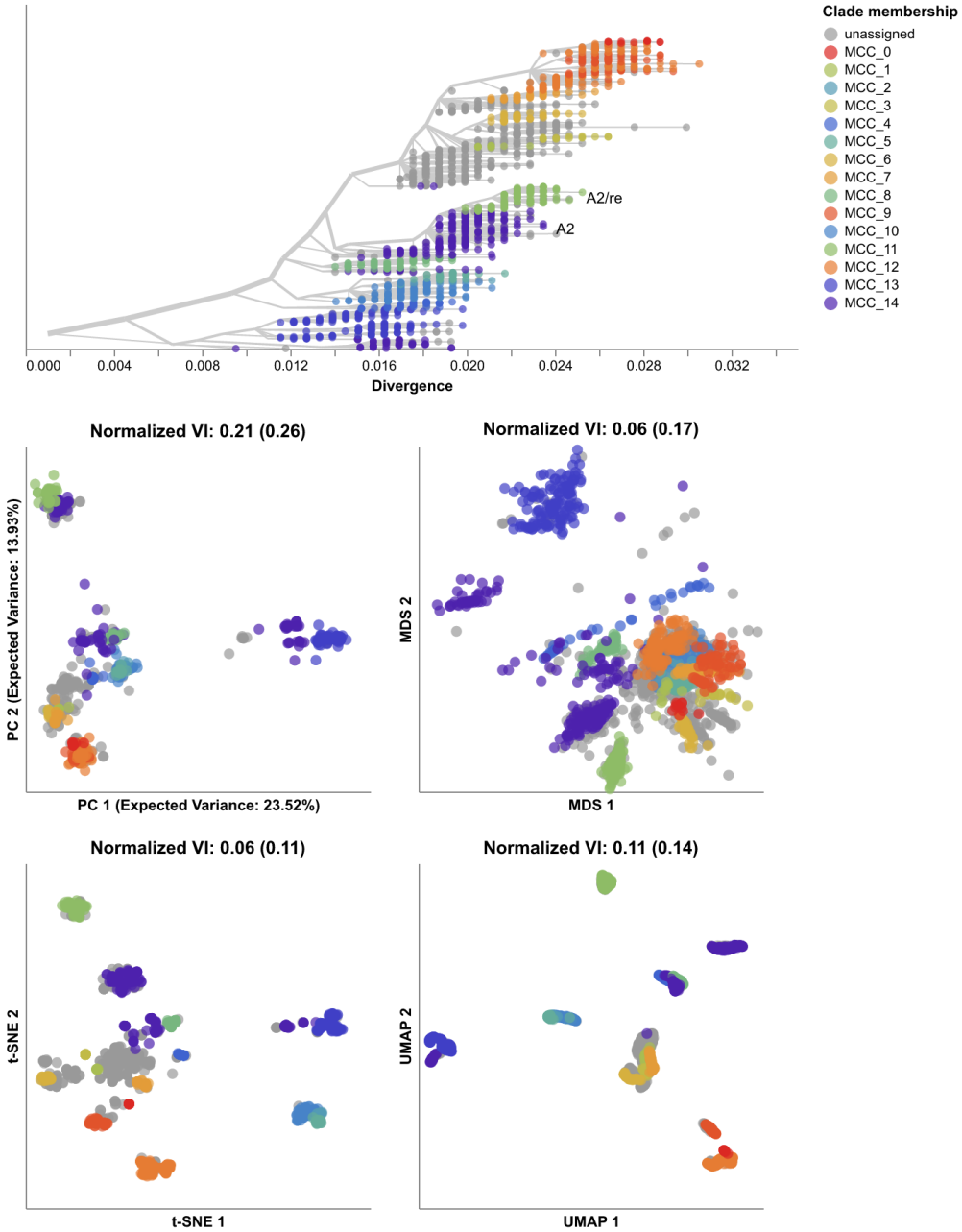

Fig 6. Phylogeny of early (2016–2018) influenza H3N2 HA sequences plotted by nucleotide substitutions per site on the x-axis (top) and low-dimensional embeddings of the same HA sequences concatenated with matching NA sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom 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. "A2" and "A2/re" labels indicate a known reassortment event [45].

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 [46], a mutation rate in the spike surface protein subunit S1 that is four times higher than influenza H3N2's HA rate [47], and increasingly common recombination [48, 49]. However, multiple expert-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.) [50] or systematically defined by the Nextstrain team [51–53] to smaller, emerging genetic clusters defined by Pango curators [19]. 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 Pango lineages (see Methods). To test 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 VI distance between those clusters and known genetic groups. Unlike the seasonal influenza analysis, we counted insertion and deletion ("indel") events in pairwise genetic distances for SARS-CoV-2, to improve the resolution of distance-based embeddings.

All embedding methods 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 21L (Omicron) placed farthest from other clades, while both Delta clades (21I and 21J) placed close together (Fig. 7, S14 Fig). MDS placed related clades closer together on a continuous scale, while PCA, t-SNE, and UMAP produced more clearly separate groups of samples. We did not observe the same clear grouping of Pango lineages. For example, the Nextstrain clade 21J (Delta) contained 11 Pango lineages that all appeared to map into the same overlapping space in all four 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 Pango nomenclature.

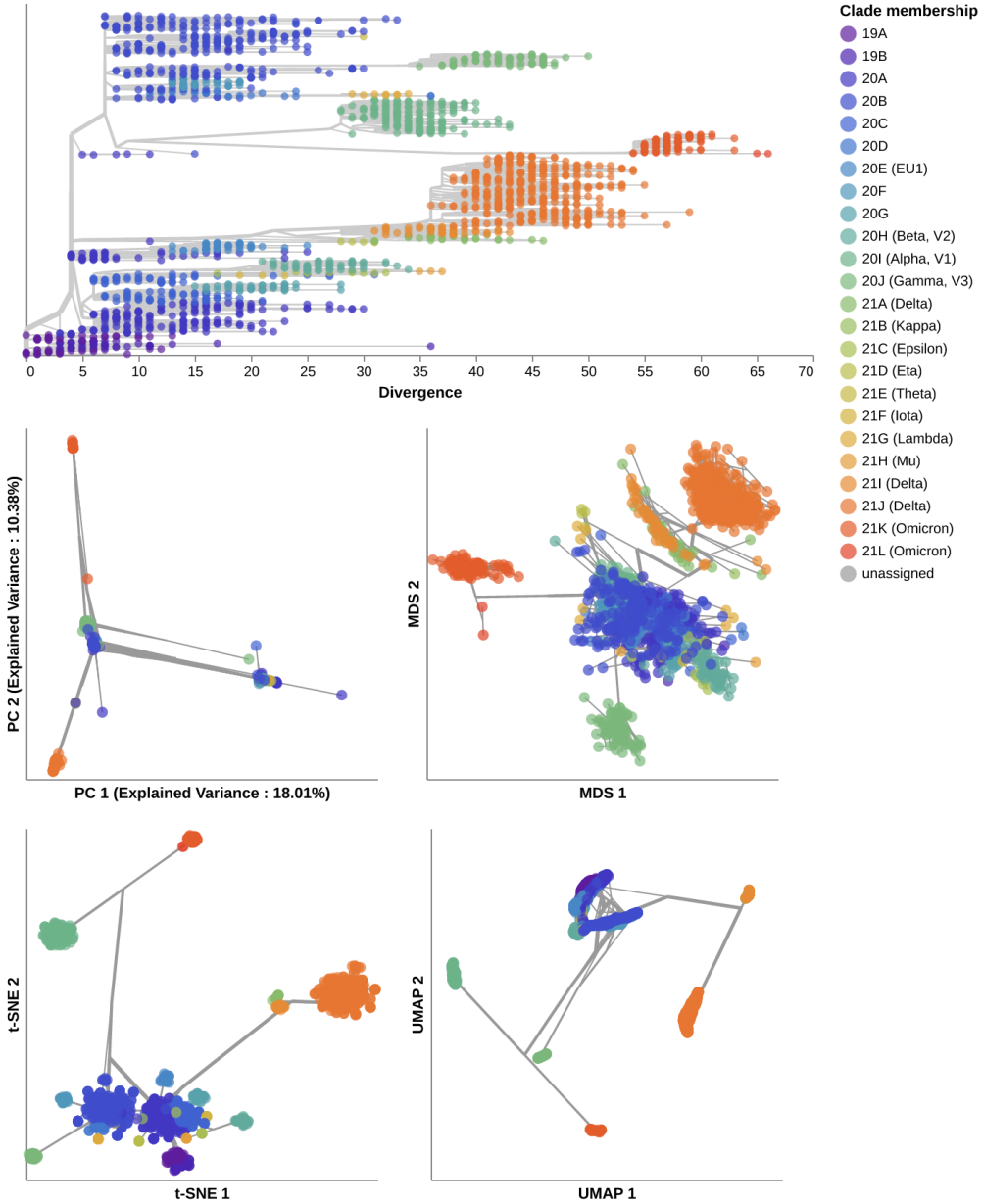

Fig 7. Phylogeny of early (2020–2022) SARS-CoV-2 sequences plotted by number of nucleotide substitutions from the most recent common ancestor on the x-axis (top) and low-dimensional embeddings of the same sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips in the tree and embeddings are colored by their Nextstrain clade assignment. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

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. PCA produced the weakest linear relationship (Pearson's $R^2 = 0.20$, Fig. 8). MDS created 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 $R^2 = 0.61$, 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 genetically similar samples in t-SNE and UMAP remained consistent up to a genetic distance of approximately 30 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 Pango lineages represented different resolutions of genetic diversity, we identified optimal distance thresholds per lineage definition. However, we found that the optimal thresholds were the same for both lineage definitions (S1 Table). Only clusters from t-SNE and UMAP represented completely monophyletic groups (S2 Table). These two methods also produced the most clusters supported by specific mutations (S3 Table). The 19 clusters from t-SNE were closest to the 24 Nextstrain clades (normalized VI=0.07), while the other methods were 2-3 times farther away (Fig. 9) Clusters from t-SNE also had the most similar within-group distances to Nextstrain clades (S16 Fig). Clusters from all methods were farther from the 35 Pango lineages (S17 Fig), but t-SNE's clusters were the closest (normalized VI=0.12). PCA, MDS, and UMAP clusters were at least twice as far from Pango lineages as t-SNE's clusters. We found that within-cluster distances for t-SNE were lower on average than within-lineage Pango distances, suggesting that Pango lineages were not as tightly scoped as we originally expected (S16 Fig). 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 defined by Pango lineage nomenclature.

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 the

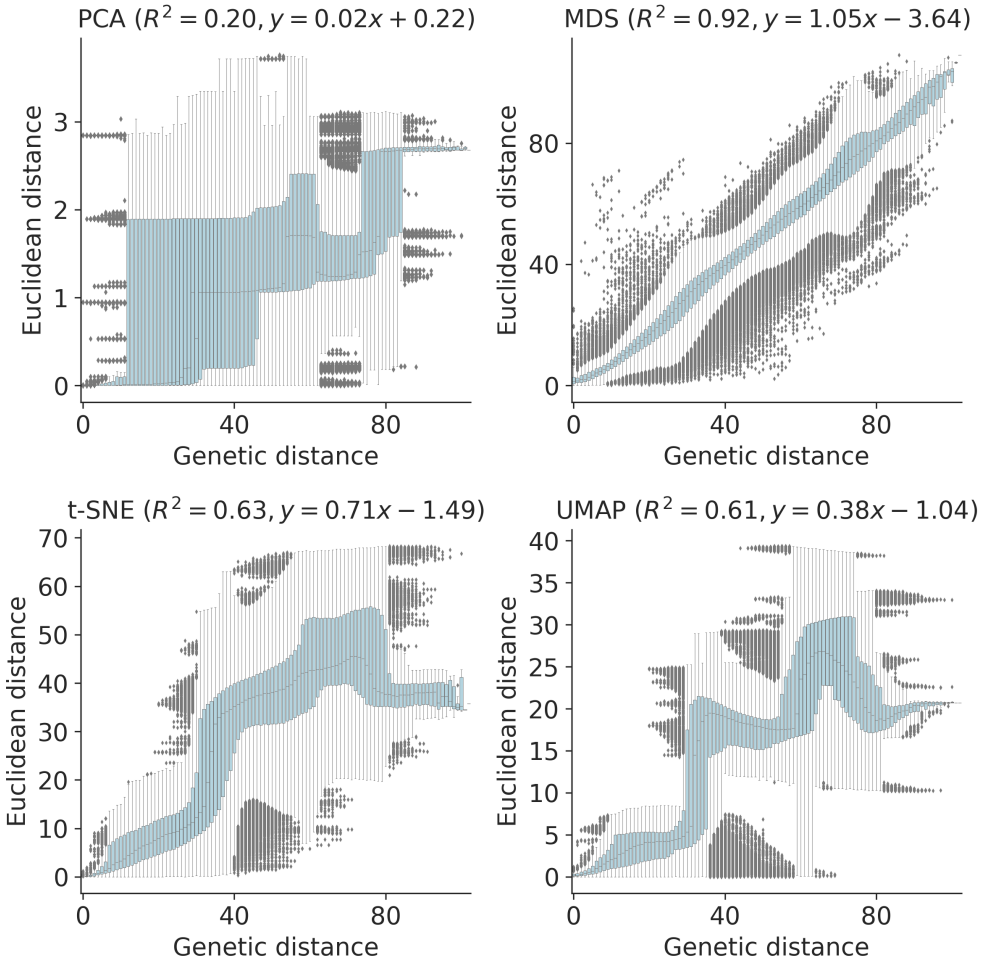

Fig 8. Relationship between pairwise genetic and Euclidean distances in embeddings for early (2020–2022) SARS-CoV-2 sequences by PCA (upper left), MDS (upper right), t-SNE (lower left), and UMAP (lower right). Each boxplot represents the distribution of pairwise Euclidean distances at a given genetic distance. Panel titles include Pearson's R^2 values and linear regression coefficients between the plotted distances.

corresponding Nextstrain clades and Pango lineages. Only t-SNE and UMAP clusters were monophyletic (S2 Table). Only PCA and t-SNE had cluster-specific mutations for more than half their clusters (S3 Table). Clusters from t-SNE had the lowest within-group distances (S16 Fig). Compared to the 18 Nextstrain clades defined in this time period, the closest clusters were from t-SNE (normalized VI=0.10) and UMAP (normalized VI=0.09, Fig. 10). However, t-SNE produced 69 clusters, over five times more than UMAP's 13. We attributed these additional clusters to distinct recombinant groups that all received the same Nextstrain clade label of "recombinant". We observed

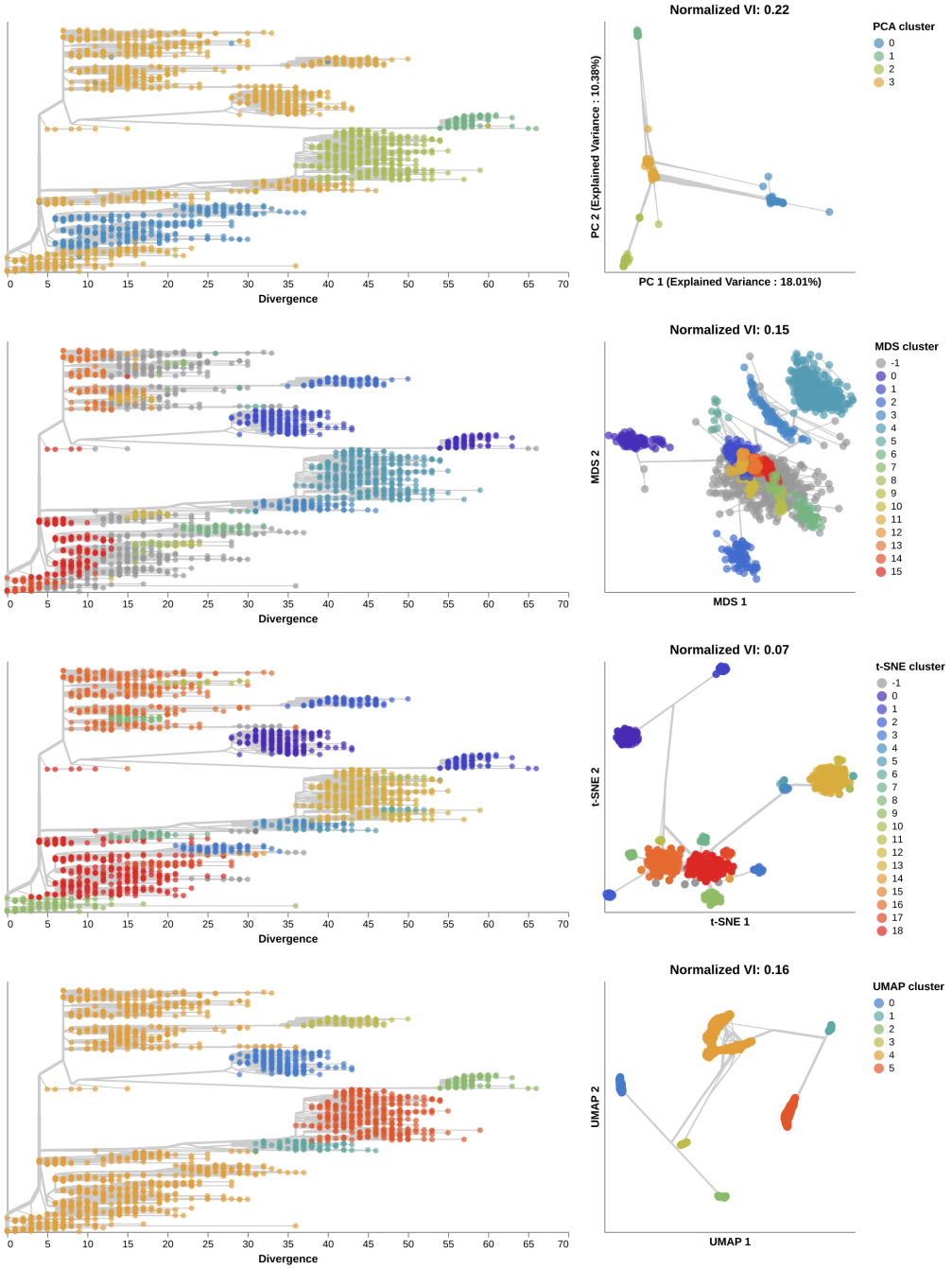

Fig 9. Phylogenetic trees (left) and embeddings (right) of early (2020–2022) SARS-CoV-2 sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Nextstrain clades). Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

similar absolute and relative distances to Nextstrain clades across methods at different sampling densities under an even geographic and temporal sampling scheme (S18 Fig). However, the presence of geographic and genetic bias associated with randomly sampling the late SARS-CoV-2 data produced less accurate t-SNE clusters and more accurate PCA, MDS, and UMAP clusters.

All methods produced less accurate representations of the 137 Pango lineages (S19 Fig). However, t-SNE clusters were nearly as accurate with a normalized VI of 0.14, suggesting that t-SNE's numerous additional clusters likely did represent many of the 69 recombinant Pango lineages in the dataset that all received a "recombinant" Nextstrain clade label. Clusters from other methods were at least twice as far from Pango lineages as t-SNE's clusters, suggesting that these other methods poorly captured recombinant lineages. With the exception of t-SNE's performance, these results replicate the patterns we observed with early SARS-CoV-2 data where clusters from embeddings more effectively represented broader genetic diversity than the finer resolution diversity denoted by Pango lineages. Unlike the 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).

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

Finally, we tested the ability of sequence embeddings to place known recombinant lineages of SARS-CoV-2 between their parental lineages in Euclidean space. We reasoned that each 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.

We identified 66 recombinant lineages for which that lineage and both of its parental lineages had at least 10 genomes (S4 Table). MDS embeddings most consistently placed

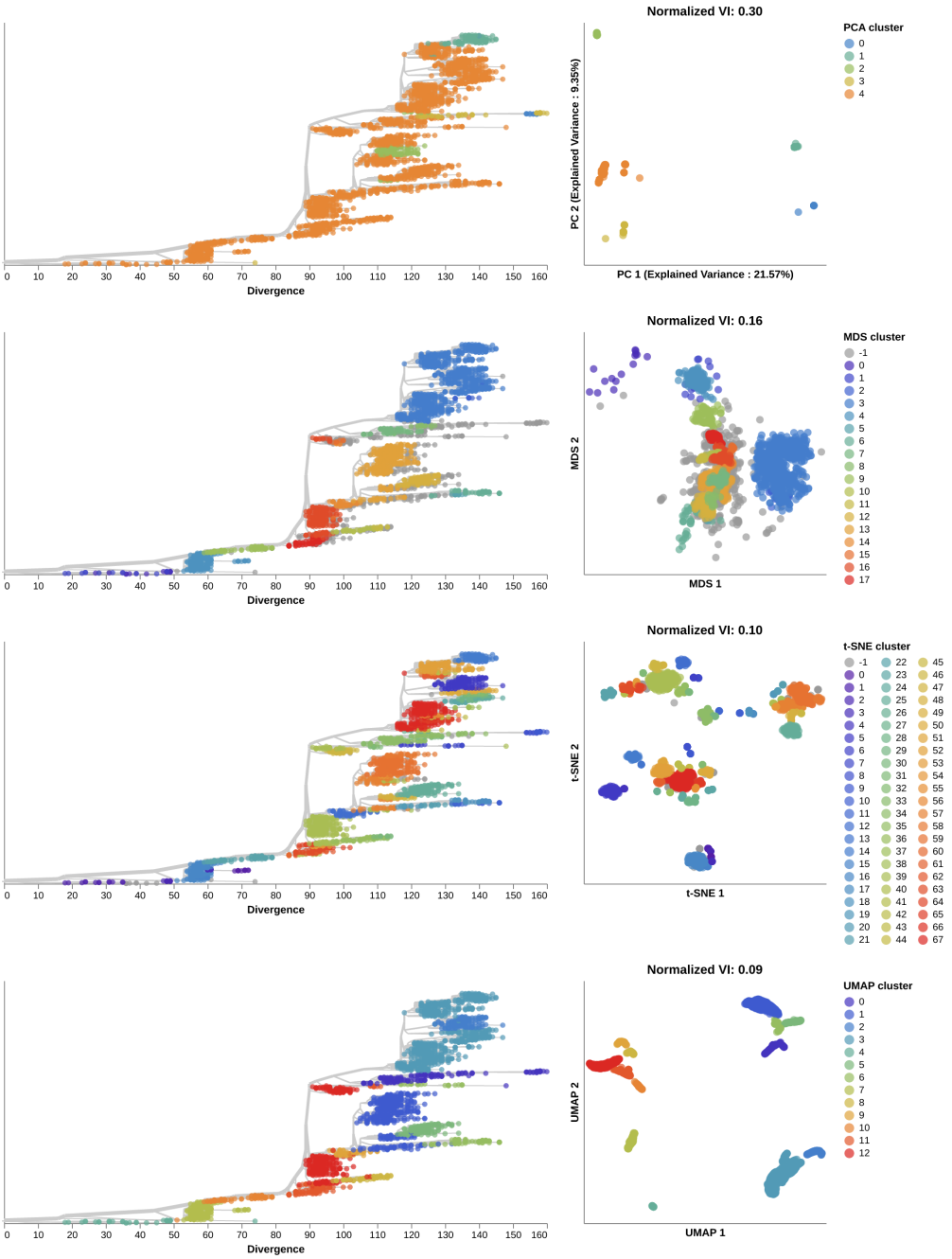

Fig 10. Phylogenetic trees (left) and embeddings (right) of late (2022–2023) SARS-CoV-2 sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Nextstrain clades).

recombinant lineages between the parental lineages with correct placement of 60 lineages (91%). The t-SNE, UMAP, and PCA embeddings correctly placed 54 (82%), 52 (79%), and 40 (61%) lineages, respectively. Additionally, all 66 recombinant lineages

placed closer to at least one parent in all embeddings except for two lineages in the PCA embeddings.

Discussion

We applied four standard dimensionality reduction methods to simulated and natural genome sequences of two relevant human pathogenic viruses and found that the 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 embeddings of genetic sequences for two distinct pathogens, more than 55 years of evolution, and a wide range of practical 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. We found that clusters from t-SNE embeddings most accurately recapitulated previously defined genetic groups at the resolution of WHO variants and Nextstrain clades and consistently produced clusters that corresponded to monophyletic groups in phylogenies. 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, while t-SNE clusters most accurately captured recombinant lineages. These results show 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 use, including indels, reassortment, and recombination.

From these results, we can also make the following recommendations about how to apply these methods to other viral pathogens. First, evenly sample the available genome sequences across time and geography, to minimize bias in embeddings. Then, choose which embedding method to use based on the question under investigation. For analyses that require the most accurate low-dimensional representation of pairwise

genetic distances across local and global scales, use MDS with 3 dimensions. For analyses that need to find clusters of closely related samples, use t-SNE with a perplexity of 100 (or less, if using fewer than 100 samples) and a learning rate that scales with the number of samples in the data. In all cases, plot the relationship between pairwise genetic distances and Euclidean distances in each embedding. These plots reveal the range of genetic distances that an embedding can represent linearly and act as a sanity check akin to plotting the temporal signal present in samples prior to inferring a time-scaled phylogeny [4, 54]. Before finding clusters in the t-SNE embedding, determine the minimum genetic distance desired between clusters, and use the pairwise genetic and Euclidean distance plot to find the corresponding Euclidean distance to use as a threshold for HDBSCAN. While HDBSCAN clusters require this pathogen-specific tuning, the linear relationship between Euclidean and genetic distance remains robust to changes in method parameters.

Despite the promise of these simple methods to answer important public health questions about human pathogenic viruses, these methods and our analyses suffer from inherent limitations. The lack of an underlying biological model is both a strength and the clearest limitation of the dimensionality reduction methods we considered here. For example, embeddings of SARS-CoV-2 genomes cannot capture the same fine-grained genetic resolution as Pango lineage annotations. Each method provides only a few parameters to tune its embeddings and these parameters have little effect on the qualitative outcome. Each method also suffers from specific issues explored in our analyses. PCA performs poorly with missing data and requires researchers to either ignore columns with missing values or impute the missing values prior to analysis, as previously shown for Zika virus [30]. Neither t-SNE nor UMAP maintain a linear relationship between pairwise Euclidean and genetic distances across the observed range of genetic distances. As a result, viewers cannot know that samples mapping far apart in a t-SNE or UMAP embedding are as genetically distant as they appear. In maintaining a linear relationship between Euclidean and genetic distances, MDS sacrifices the ability to form more accurate genetic clusters for viruses with large genomes like SARS-CoV-2. Given these limitations of these methods, we do not expect them to replace biologically-informed methods that provide more meaningful parameters to tune their algorithms. Instead, these methods provide an easy first step

to produce interpretable visualizations and clusters of genome sequences, prior to analysis with more sophisticated methods with biological models.

We note that our analysis reflects a small subset of human pathogenic viruses and dimensionality reduction methods. We focused on analysis of two respiratory RNA viruses that contribute substantially to seasonal human morbidity and mortality, but numerous alternative pathogens would also have been relevant subjects. For example, HIV represents a canonical example of a highly recombinant and bloodborne virus, while Zika, dengue, and West Nile viruses represent pathogens with multiple host species in a transmission chain. Similarly, we selected only four dimensionality reduction methods from myriad options that are commonly applied to genetic data [55]. We chose these methods based on their wide use and availability in tools like scikit-learn [56] and to limit the dimensionality of our analyses.

Some limitations noted above suggest future directions for this line of research. We provide optimal settings for each pathogen and embedding method in this study and open source tools to apply these methods to other pathogens. Researchers can easily integrate these tools into existing workflows for the genomic epidemiology of viruses and visualize the results with Nextstrain. Alternately, researchers may choose to apply similar existing tools developed for analysis of metagenomic or bacterial populations [57–61] to the analysis of viral populations. In the short term, researchers can immediately apply the methods we describe here to seasonal influenza and SARS-CoV-2 genomes to identify biologically relevant clusters. Researchers can also apply these methods to find relevant clusters for other viruses by evaluating the pairwise Euclidean and genetic distances for each virus and tuning the Euclidean distance thresholds for HDBSCAN to capture the desired granularity of genetic clusters. In the long term, we expect researchers will benefit from expanding the breadth of dimensionality reduction methods applied to viruses and the breadth of viral diversity assessed by these methods. Additionally, the combination of dimensionality reduction methods and clustering with HDBSCAN provides the foundation for future methods to automatically identify reassortant and recombinant lineages.

Conclusion

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

Materials and methods

Embedding methods

We selected four standard and common dimensionality reduction (or "embedding") methods to apply to human pathogenic viruses: PCA, MDS, t-SNE, and UMAP. PCA operates on a matrix with samples in rows, "features" in columns, and numeric values in each cell [22]. To apply PCA to multiple sequence alignments, we transformed each nucleotide value into a corresponding integer (A to 1, G to 2, C to 3, T to 4, and all other values to 5) and applied scikit-learn's PCA implementation to the resulting numerical matrix with the "full" singular value decomposition solver and 10 components [56]. To minimize the effects of missing data on the PCA embeddings, we dropped all columns with "N" or "-" characters from concatenated H3N2 HA/NA alignments and SARS-CoV-2 alignments prior to producing PCA embeddings.

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

We applied scikit-learn's MDS implementation to a given distance matrix, with an option to set the number of components in the resulting embedding [56]. Similarly, we applied scikit-learn's t-SNE implementation, with options to set the "perplexity" and

the "learning rate". The perplexity controls the number of neighbors the algorithm uses per input sample to determine an optimal embedding [24]. This parameter effectively determines the balance between maintaining "local" or "global" structure in the embedding [37]. The learning rate controls how rapidly the t-SNE algorithm converges on a specific embedding [24, 62] and should scale with the number of input samples [63]. We initialized t-SNE embeddings with the first two components of the corresponding PCA embedding, as previously recommended to obtain more accurate global structure [34, 37]. Finally, we applied the *umap-learn* Python package written by UMAP's authors, with options to set the number of "nearest neighbors" and the "minimum distance" [25]. As with t-SNE's perplexity parameter, the nearest neighbors parameter determines how many adjacent samples the UMAP algorithm considers per sample to find an optimal embedding. The minimum distance sets the lower limit for how close any two samples can map next to each other in a UMAP embedding. Lower minimum distances allow tighter groups of samples to form. For both t-SNE and UMAP, we used the default number of components of 2.

Simulation of influenza-like and coronavirus-like populations

Given the relative lack of prior application of dimensionality reduction methods to human pathogenic viruses, we first attempted to understand the behavior and optimal parameter values for these methods when applied to simulated viral populations with well-defined evolutionary parameters. To this end, we simulated populations of influenza-like and coronavirus-like viruses using SANTA-SIM [64]. These simulated populations allowed us to identify optimal parameters for each embedding method, without overfitting to the limited data available for natural viral populations. For each population type described below, we simulated five independent replicates with fixed random seeds for over 55 years, filtered out the first 10 years of each population as a burn-in period, and analyzed the remaining years.

We simulated influenza-like populations as previously described with 1,700 bp hemagglutinin sequences [65]. As in that previous study, we scaled the number of simulated generations per real year to 200 per year to match the observed mutation rate for natural H3N2 HA sequences, and we sampled 10 genomes every 4 generations for

12,000 generations (or 60 years of real time).

We simulated coronavirus-like populations as previously described for human seasonal coronaviruses with genomes of 21,285 bp [12]. For the current study, we assigned 30 generations per real year to obtain mutation rates similar to the 8×10^{-4} substitutions per site per year estimated for SARS-CoV-2 [66]. To account for the effect 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, 67]. 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 [68]. 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 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, 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 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 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 predefined error threshold [23].

With the approach described above, we tested each method across a range of relevant parameters with all combinations of parameter values. For PCA and MDS, we tested the number of components between 2 and 10. 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.

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 [41]. 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.

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

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

For SARS-CoV-2 data, we defined the early dataset between January 1, 2020 and January 1, 2022 and the late dataset between January 1, 2022 and November 3, 2023. For the early dataset, we evenly sampled 1,736 SARS-CoV-2 genomes by geographic region, year, and month, excluding known outliers. For the late dataset, we used the same even sampling by space and time to select 1,309 representative genomes. In addition to these genomes, we identified all recombinant lineages in the official Pango designations as of November 3, 2023 (`https://github.com/cov-lineages/pango-designation/raw/1bf4123/pango_designation/alias_key.json`) for which the recombinant lineage and both of its parental lineages had at least 10 genome records each. We sampled at most 10 genomes per lineage for all distinct recombinant and parental lineages for a total of 1,157. With these additional genomes, the late SARS-CoV-2 dataset included 2,464 total genomes.

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

To evaluate the biological interpretability of distances between samples in low-dimensional embeddings, we plotted the pairwise Euclidean distance between samples in each embedding against the corresponding genetic distance between the same samples. We calculated Euclidean distance using all components of the given embedding (e.g., 2 components for PCA, t-SNE, and UMAP and 3 components for MDS). For each embedding, we fit a linear model between Euclidean and genetic distance and calculated the squared Pearson's correlation coefficient, R^2. The distance plots provide a qualitative assessment of each embedding's local and global structure relative to a 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 phylogenetic tree. For seasonal influenza H3N2 HA and NA sequences, we aligned sequences with MAAFT (version 7.486) [69, 70] using the *augur align* command (version 22.0.3) [71]. 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) and generated with Nextalign (version 2.14.0) [21]. We inferred a phylogeny with IQ-TREE (version 2.1.4-beta) [72] using the *augur tree* command and named internal nodes of the resulting divergence tree with TreeTime (version 0.10.1) [4] using the *augur refine* command. We visualized phylogenies with Auspice [73], after first converting the trees to Auspice JSON format with *augur export*. To visualize phylogenetic relationships in the context of each pathogen embedding, we calculated the mean Euclidean position of each internal node in each dimension of a given embedding (e.g., MDS 1) based on the Euclidean positions of that node's immediate descendants and plotted line segments on the embedding connecting each node of the tree with its immediate parent to represent branches in the phylogeny. We only plotted these phylogenetic relationships on embeddings for pathogen datasets that lacked reassortment and recombination including early and late H3N2 HA and early SARS-CoV-2 datasets.

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 [42].

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 (version 0.5.6) [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 of these TreeKnit expectations, we inferred HA and NA trees with IQ-TREE with a custom argument to collapse near-zero-length branches ('-czb'). We rooted the resulting trees on the same sample that we used as an alignment reference, A/Beijing/32/1992, and pruned this sample prior to downstream analyses. We applied TreeKnit to the rooted HA and NA trees with a gamma value of 2.0 and the '–better-MCCs' flag, as previously recommended for H3N2 analyses [11]. Finally, we filtered the MCCs identified by TreeKnit to retain only those with at least 10 samples and to omit the root MCC that represented the most recent common ancestor in both HA and NA trees.

For SARS-CoV-2, we used both coarser "Nextstrain clades" [51–53] and more granular Pango lineages [19] provided by Nextclade as "Nextclade pango" annotations. Nextstrain clade definitions represent the World Health Organization's variants of concern along with post-Omicron phylogenetic clades that have reached minimum global and regional frequencies and growth rates. Pango lineages represent expert-curated lineages (https://github.com/cov-lineages/pango-designation) and must contain at least 5 samples with an unambiguous evolutionary event. Additionally, Pango lineages produced by recombination receive a lineage name prefixed by an "X", while Nextstrain clades do not explicitly reflect recombination events.

Since Pango lineages can represent much smaller genetic groups than are practically useful, we collapsed lineages with fewer than 10 samples in our analysis into their parental lineages using the pango_aliasor tool (https://github.com/corneliusroemer/pango_aliasor). Specifically, we counted the number of samples per lineage, sorted lineages in ascending order by count, and collapsed each lineage with a count less than 10 into its parental lineage in the count-sorted order. This approach allowed small lineages to aggregate with other small

parental lineages and meet the 10-sample threshold. We used these "collapsed Nextclade Pango" lineages for subsequent analyses.

Clustering of samples in embeddings

To understand how well embeddings of genetic data could capture previously defined genetic groups, we applied an unsupervised clustering algorithm, HDBSCAN [43], 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 variation in embedding-specific distances, we performed a coarse grid search of distance threshold values for each virus type and embedding method.

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 [44]. 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 "Pango lineage" for SARS-CoV-2). HDBSCAN allows samples to not belong to a cluster and assigns these samples a numeric label of -1. We intentionally included all unassigned samples in the normalized VI calculation thereby penalizing cluster parameters that increased the number of unassigned samples by increasing their VI values. Since Nextstrain clade assignments could include non-monophyletic labels like "unassigned" and "recombinant" to represent samples that did not map into a single clade, we ignored these labels in our VI distance calculations to avoid rewarding cluster that placed such non-monophyletic samples into the same group. Finally, we used these optimal distance thresholds to identify clusters in out-of-sample data from the late datasets for both viruses and calculate the normalized VI between those clusters and previously defined genetic groups.

Evaluating robustness of embedding cluster accuracy

The cluster accuracies we estimated for late H3N2 HA and SARS-CoV-2 datasets represented a single VI measurement for a single pathogen dataset. To understand how robust these accuracies were across different datasets, we generated alternate random samples from both late pathogen datasets using two different sampling schemes and a range of total sequences sampled. Specifically, we sampled 500, 1000, 1500, 2000, or 2500 total sequences for five replicates per pathogen (random seeds of 0, 1, 2, 3, and 4) with either even sampling by geography and time or random sampling. For the relatively smaller influenza data, we evenly sampled by country, year, and month. For the larger SARS-CoV-2 data, we evenly sampled by region, year, and month. Even sampling attempted to minimize geographic and temporal biases in the original data. Random sampling uniformly selected samples in a way that reflected the bias in the data. Influenza data were heavily biased toward samples from the USA and clade 3c3.A, while SARS-CoV-2 data were biased toward Europe and North America and Nextstrain clades 21K, 21L, and 22B. For each replicate from each sampling scheme and total number of sequences, we embedded the corresponding sequences with each method, identified clusters in embeddings, and calculated the VI distance between those clusters

and Nextstrain clade assignments. We plotted the distribution of the resulting VI distances, to estimate the variance of these values caused by sampling bias and density.

Evaluating the monophyletic nature of embedding clusters

To quantify the degree to which embedding clusters represented monophyletic groups in a pathogen phylogeny, we counted the number of times clusters from each embedding method appeared in different parts of the tree. Specifically, we applied *augur traits* with TreeTime (version 0.10.1) [4, 71] to infer cluster labels for internal nodes of the phylogeny for each pathogen dataset and embedding method. Using a preorder traversal of the tree, we identified each transition between different cluster labels assigned to pairs of ancestral and derived internal nodes. Since the "unclustered" cluster label of "-1" produced by HBSCAN could occur in both ancestral and derived nodes and lead to overcounting transitions, we only logged transitions with this label in the ancestral state (e.g., transition from cluster -1 to cluster 0 but not cluster 0 to cluster -1). For each embedding, we counted the number of distinct clusters, total transitions, and excess transitions beyond the expected single transition between pairs of clusters. Embeddings with no excess transitions between clusters represented monophyletic groups.

Identification of cluster-specific mutations

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

Assessment of HA/NA reassortment in seasonal influenza populations

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. As mentioned above, we dropped all columns with "N" or "-" 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.

Assessment of recombination in SARS-CoV-2 populations

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 Pango annotations to identify samples in each lineage, as these were the lineage names used to include recombinant samples in the analysis and define known relationships between recombinant and parental lineages.

Data and software availability

The entire workflow for our analyses was implemented with Snakemake [74]. We have provided all source code, configuration files, and datasets at https://github.com/blab/cartography. Interactive phylogenetic trees and corresponding embeddings for natural populations are available at https://nextstrain.org/groups/blab/ under the "cartography" keyword. The *pathogen-embed* Python package, available at https://pypi.org/project/pathogen-embed/, provides command line utilities to calculate distance matrices (*pathogen-distance*), calculate embeddings per method (*pathogen-embed*), and apply hierarchical clustering to embeddings (*pathogen-cluster*).

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 specific embedding method (PCA, MDS, t-SNE, and UMAP) and a given combination of method parameters. Boxplots reflect median, upper and lower quartiles, and the range of values.

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 specific embedding method (PCA, MDS, t-SNE, and UMAP) and a given combination of method parameters. Boxplots reflect median, upper and lower quartiles, and the range of values.

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

S4 Fig. MDS embeddings for early (2016–2018) influenza H3N2 HA sequences showing all three components. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line colors

represent the clade membership of the most ancestral node in the pair of nodes connected by the segment. Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

S5 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.

S6 Fig. Phylogeny of late (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. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line colors represent the clade membership of the most ancestral node in the pair of nodes connected by the segment. Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

S7 Fig. MDS embeddings for late (2018–2020) influenza H3N2 HA sequences showing all three components. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line colors represent the clade membership of the most ancestral node in the pair of nodes connected by the segment. Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

S8 Fig. Replication of cluster accuracy per embedding method for late (2018–2020) influenza H3N2 HA sequences across different sampling densities (total sequences sampled) and sampling schemes including A) even geographic and temporal sampling and B) random sampling. We measured cluster accuracy across five replicates per sampling density and scheme with the

normalized VI distance between clusters from a given embedding and Nextstrain clades for the same samples. The even sampling scheme selected sequences evenly across country, year, and month to minimize geographic and temporal bias. The random sampling scheme uniformly sampled from the original dataset, reflecting the geographic and genetic bias in those data.

S9 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).

S10 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).

S11 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).

S12 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).

S13 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).

S14 Fig. MDS embeddings for early SARS-CoV-2 sequences showing all three components. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their

immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

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 Pango lineage assignment. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

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, Pango lineages, and clusters from PCA, MDS, t-SNE, and UMAP embeddings.

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 (Pango lineages). Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.

S18 Fig. Replication of cluster accuracy per embedding method for late (2022–2023) SARS-CoV-2 sequences across different sampling densities (total sequences sampled) and sampling schemes including A) even geographic and temporal sampling and B) random sampling. We measured cluster accuracy across five replicates per sampling density and scheme with the normalized VI distance between clusters from a given embedding and Nextstrain clades for the same samples. The even sampling scheme selected sequences evenly across

region, year, and month to minimize geographic and temporal bias. The random sampling scheme uniformly sampled from the original dataset, reflecting the geographic and genetic bias in those data.

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 (Pango lineages).

S1 Table. Optimal cluster thresholds per pathogen, known genetic group type, and embedding method based on normalized variation of information (VI) distances calculated from early pathogen datasets. Smaller VI values indicate fewer differences between HDBSCAN clusters and known genetic groups. VI of 0 indicates identical clusters and 1 indicates maximally different clusters. Threshold refers to the minimum Euclidean distance between initial clusters for HDBSCAN to consider them as distinct clusters. We apply these optimal thresholds per pathogen, known genetic group type, and method to find clusters in corresponding late datasets for each pathogen.

S2 Table. Number of clusters, transitions between clusters in the phylogeny, and excess transitions indicating non-monophyletic groups per pathogen and embedding. Embeddings without any excess transitions reflect monophyletic groups in the corresponding pathogen phylogeny.

S3 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%. Cluster-specific mutations appear in rows with a *cluster_count* value of 1.

S4 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.

S5 Table. Accessions and authors from originating and submitting laboratories of seasonal influenza and SARS-CoV-2 sequences from INSDC databases.

Acknowledgments

We thank James Hadfield, Katie Kistler, Maya Lewinsohn, Nicola Muller, Louise Moncla, Nidia Trovao, and Michael Zeller for constructive feedback on this project. We gratefully acknowledge the originating and submitting laboratories of seasonal influenza and SARS-CoV-2 sequences from INSDC databases without whom this work would not be possible (S5 Table).

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 Comput Biol. 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. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Phil Trans R Soc A. 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. arXiv. 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. He J, Deem MW. Low-dimensional clustering detects incipient dominant influenza strain clusters. Protein Eng Des Sel. 2010;23(12):935–946.

33. Ito K, Igarashi M, Miyazaki Y, Murakami T, Iida S, Kida H, et al. Gnarled-trunk evolutionary model of influenza A virus hemagglutinin. PLoS One. 2011;6(10):e25953.

34. 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.

35. Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS Genet. 2019;.

36. 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;.

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

38. Chari T, Pachter L. The specious art of single-cell genomics. PLoS Comput Biol. 2023;19(8):e1011288.

39. Petrova VN, Russell CA. The evolution of seasonal influenza viruses. Nature Reviews Microbiology. 2018;16(1):47–60. doi:10.1038/nrmicro.2017.118.

40. 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.

41. Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2021;49(D1):D121–D124.

42. Neher RA, Bedford T. nextflu: real-time tracking of seasonal influenza virus evolution in humans. Bioinformatics. 2015;31(21):3546–3548.

43. 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.

44. 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.

45. 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.

46. 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.

47. 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.

48. Focosi D, Maggi F. Recombination in coronaviruses, with a focus on SARS-CoV-2. Viruses. 2022;14(6).

49. 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.

50. 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.

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. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;2(1):vew007.

55. 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.

56. 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.

57. 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.

58. Schloss PD. Reintroducing mothur: 10 Years Later. Appl Environ Microbiol. 2020;86(2).

59. 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.

60. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8(4):e61217.

61. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019;29(2):304–316.

62. 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.

63. 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.

64. 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).

65. 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.

66. Rambaut A. Phylogenetic analysis of nCoV-2019 genomes; 2020. `https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356`.

67. 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.

68. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 3rd ed. Melbourne, Australia: OTexts; 2021. Available from: `OTexts.com/fpp3`.

69. 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.

70. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780.

71. 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).

72. 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.

73. 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.

74. 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.