Genetic cartography reveals ancestral relationships of human pathogenic viruses

Sravani Nanduri[1,2]

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

Abstract

Phylogenetics is vital to epidemiologists' understanding of population patterns, dynamics, and transmission, and is integral to public health studies. Most diseases can be modeled in a transmission tree, an approach that tracks mutations in disease samples back to a single common ancestor; however, issues with sample quality, recombination, and other factors can make it difficult to quantify a genetic sequence's mutations in reference to others. In this paper, we study the use of dimensionality reducing algorithms PCA, MDS, t-SNE, and UMAP in understanding viral population structure, and present quantitative and qualitative interactive visualizations that summarize the accuracy and scope of these models. With the public tools and automated code pipelines for ease of integration into other research projects, this paper will discuss the practical applications and future steps for this form of analysis and modeling within the scope of viral epidemiology.

Introduction

Phylogenetic inference is a fundamental tool for understanding genealogical relationships among human pathogenic viruses. However, recombination and reassortment in viral populations, the process of multiple viruses invading the same cell, which during cloning causes the viruses to exchange genetic material between each other, invalidates basic phylogenetic assumptions of inheritance and requires more sophisticated approaches (Pérez-Losada et al. 2015). One approach is to split a genome into multiple phylogenies to model the evolution of the nonrecombinant fragments. This is done using a genetic algorithm that scans strains for recombination breakpoints, quantifies and analyzes the impact of recombination at each one, and splits the phylogeny at its most important breakpoints (Kosakovsky Pond et al. 2006). Finding recombination breakpoints relies on the detection of a recombination signal through methods such as CHIMAERA and LARD. Both CHIMAERA and LARD use split decomposition, a method which depicts parallel edges between sequences if there are conflicting phylogenetic signals (Posada and Crandall 2001; Martin et al. 2017).

An alternate strategy is to compare viral genomes with methods that do not make the same strong assumptions as phylogenetic inference. Principal component analysis (PCA) has been used to visualize human population structure from genomic variants (Novembre et al. 2008; Alexander et al. 2009; Peter H. Sudmant 2015) PCA was also used to reveal Zika's genetic diversity and spread in the Americas by assessing the clustering of multidimensional genetic data (H. C. et al. 2017). Principal component analysis (PCA) was consistent with the phylogenetic observations, and showed tight clustering in Zika genomes in strains from the same geographical introduction. MDS has been applied to H3N2 sequences to inspect relationships between all gene segments, which is closely related to the subject of this paper, with the difference that Rambaut et al. (2008) looks at between-gene diversity rather than within-gene. The MDS analyses showed tight clustering between genes, suggesting that the evolutionary dynamics of influenza A virus is shaped to some degree by phylogenetic history and global epidemiological dynamics. PCA, t-SNE, and UMAP have all been used to capture both discrete and continuous patterns of variation in human genomes across a genetic continuum, and the embeddings revealed relationships between genotype, phenotype, and geography (Diaz-Papkovich et

al. 2019). While Diaz-Papkovich et al. (2019) and H. C. et al. (2017) explored qualitative measurements of embedding accuracy and fitness, this paper will go beyond that by establishing quantitative measurements for the fit and accuracy of the embeddings to further bridge the gap between visualization and statistical testing. This paper will also give insight into different reduction techniques, and will discuss both their limitations and strengths in the realm of viral data.

We present a novel approach to understanding relationships among viral genomes by transforming genomic data and then using dimensionality reduction methods such as PCA, MDS, t-SNE, and UMAP. We use interactive visualizations of the embeddings for a deeper exploration of patterns between the embeddings and the phylogeny. We investigate the degree to which this method can recapitulate known phylogenetic relationships for viruses whose genomes are phylogenetically tractable, as well as the scope of this approach when considering the intrinsic variety in clade definition and viral transmission (influenza H3N2 HA and Zika). We apply this method to viruses with less samples, lower quality strains, and genomes known to undergo substantial recombination (MERS) to assess how well each method is able to reconstruct previously identified biologically-meaningful clusters.

Results

Table 1: [All KDE values for each virus. MCC (Matthews Correlation Coefficient), FN (False Negative) FP (False Positive) TN (True Negative) TP (True Positive) accuracy(accuracy of the confusion matrix, TN + TP / total samples) threshold (z-score value the Support Vector Machine picked that best separates within vs between clade relationships) embedding (pca, mds, t-sne, or umap) between and within (z score values of the median of between clade and within clade KDE density curves, respectively)]

	MCC	FN	FP	TN	TP	accuracy	embedding	between	within	threshold
Influenza	0.576	11237	69685	516435	63318	0.878	mds	0.104	-1.476	-0.868
	0.604	11167	61032	525088	63388	0.891	genetic	0.11	-1.418	-0.88
	0.664	4796	60044	526076	69759	0.902	pca	0.28	-1.659	-1.138
	0.67	1707	66501	519619	72848	0.897	umap	0.172	-1.481	-1.048
	0.764	4268	34749	551371	70287	0.941	t-sne	0.159	-1.666	-1.09
Zika	0.663	3379	21376	172773	29273	0.891	pca	0.07	-1.434	-0.868
	0.42	6450	47231	146918	26202	0.763	mds	0.083	-1.088	-0.622
	0.54	3644	36918	157231	29008	0.821	t-sne	-0.201	-1.094	-0.736
	0.378	7830	49454	144695	24822	0.747	umap	-0.551	-0.786	-0.67
	0.513	5608	34603	159546	27044	0.823	genetic	-0.063	-1.128	-0.694
MERS	0	1388	0	32803	0	0.959	pca	-0.489	-0.504	nan
	0.625	234	1163	31640	1154	0.959	mds	-0.117	-1.627	-1.27

MCC	FN	FP	TN	TP	accuracy	embedding	between	within	threshold
0.626	135	1466	31337	1253	0.953	t-sne	0.02	-1.526	-1.372
0.479	309	2164	30639	1079	0.928	umap	0.087	-1.348	-1.234
0.712	197	754	32049	1191	0.972	genetic	0.135	-1.865	-1.42

Expectations for PCA, MDS, t-SNE, and UMAP

We have used four different methods commonly used in viral epidemiology for reducing our data, picked because of their varying ways of displaying data. Principal Component Analysis (PCA) reduces multidimensional data, increasing interpretability while minimizing information loss (Jolliffe and Cadima 2016). PCA relies on linear assumptions and does not affect the scale of the data. Because PCA is almost entirely focused on retaining the global structure and variance of the data, one of its limitations is revealing patterns locally. PCA is a matrix analysis method, while the other three methods reduce distance based comparisons. Multidimensional Scaling (MDS) refers to statistical techniques that increase the interpretability of local relational structures mired in the dataset (Hout et al. 2012). The MDS algorithm places a higher importance on translating dissimilarity to distance than displaying local patterns. t-distributed Stochastic Neighbor Embedding (t-SNE) projects clusters and distances between clusters that are not analogous to dissimilarity - in other words, t-SNE focuses more heavily on projecting similarity rather than dissimilarity (Maaten and Hinton 2008). Because t-SNE reduces dimensionality based on the local properties of data, data with intrinsically high dimensional structure will not be projected accurately. Uniform Manifold Approximation and Projection (UMAP) is a manifold learning technique for dimension reduction (McInnes et al. 2018). UMAP results in low dimensional neighborhoods that group genetically similar strains together on a local scale while still preserving relationships between distantly related strains. A limitation for UMAP is its novelty, as there are no firmly established practices and robust libraries to aid users.

Expectations for influenza, Zika, and MERS

We selected 3 representative viruses, selected for genome length, nucleotide diversity within the population, and levels of complexity to analyze their populations for this project.

H3N2 influenza is used as a proof of concept, as H3N2 HA influenza's sequences are biologically relevant, short, relatively diverse (nucleotide diversity pi = 0.0149), and only reassort and do not recombine. H3N2 influenza is a seasonal, global disease where clades are defined by phylogenetic distance (amount of mutations) from other strains. H3N2's Hemagglutinin sequences were subset from the full genome to analyze, as the HA sequences have a relatively high mutation rate compared to the other gene segments, encodes a protein that is a target of human immunity, and has traditionally been used for analysis of influenza evolution. The genomes are usually 1701 bases long, with a mean bases missing of 0.0452 and median of 0.

Zika has a longer genome, lower diversity (pi = 0.00535), a significant amount of missing bases, and can recombine. While H3N2 influenza is a globally distributed virus that has caused infections seasonally for decades, Zika is a fairly new human pathogenic virus that has a restricted geographic distribution that recapitulates the patterns of viral transmission. Therefore, Zika's clades were defined by significant geographical introductions and outbreaks. Because of the difference in clade definition from influenza, we used Zika to determine if the embeddings can recapitulate geographically significant clusters. The genomes are 10769 bases long, with a mean bases missing of 913.613 and median of 154.

With a much longer genome, homologous population (pi = 0.00235), recombination, many missing bases, and multiple hosts, MERS tests and challenges the embeddings' abilities to reveal population patterns. MERS is a recombinant virus that affects both camels and humans, with camel to human transmissions creating a need for a multiple host phylogeny. While influenza's clades are defined by mutations and Zika's by significant geographical introductions, MERS clades were assigned to internal nodes and tips in the tree based on

monophyletic host status (strictly camel or human) to reveal patterns within host outbreaks. The genomes are 30130 bases long, with a mean bases missing of 889.781 and median of 42.5. MERS population is relatively homologous, as the majority of human infections were collected from the same outbreaks.

Embedding clusters recapitulate phylogenetic clades for seasonal influenza A/H3N2

All four dimensionality reduction methods qualitatively recapitulated clade-level groupings observed in the phylogeny (Figure 1). Strains from the same clade appeared tightly grouped in PCA, t-SNE, and UMAP embeddings and more loosely clustered in the MDS embedding. Closely related clades tended to tightly cluster in PCA, MDS, UMAP, and, to a lesser extent, t-SNE. For example, the clade A2 and its subclade A2/re map to adjacent regions of all four embeddings. We observed the same pattern for A1 and its subclade A1a as well as for A1b and its subclades A1b/135K and A1b/135N. The clade 3c2.A and its subclade A3 clustered in all embeddings except t-SNE. This result matched our expectation that t-SNE would preserve local clusters and not retain global structure between more distantly related data.

To quantify the patterns we observed in Figure 1, we calculated two complementary metrics for each embedding method. First, we measured the linearity of the relationship of Euclidean distance between two strains in an embedding space and the genetic distance between these same strains. All four methods exhibited a consistent linear relationship for pairs of strains that differed by no more than 30 nucleotides (Figure 2). PCA and MDS provided the strongest linear mapping to genetic distance (Pearson's $R^2 = 0.69 \pm 0.001$ and 0.68 ± 0.001, respectively). This same mapping for the UMAP method was less of a linear function (Pearson's $R^2 = 0.38 \pm 0.001$) than a piecewise function of two parts. Strain pairs with more than 30 nucleotide differences were not as well separated in UMAP space as strains with lesser genetic distances. This result suggests that UMAP might be most effective for distinguishing between more distantly related strain pairs. t-SNE's mapping was the weakest (Pearson's $R^2 \pm 0.001$) and revealed that only closely related strains map near each other in t-SNE space. Pairs of strains that differ by more than 15 nucleotides are unlikely to be placed near each other in a t-SNE embedding.

Second, we determined how accurately the Euclidean distance between pairs of strains in an embedding could classify those strains as belonging to the same clade or not. Specifically, we used a support vector machine (SVM) classifier to identify an optimal Euclidean distance threshold that distinguished pairs of strains from the same clade. To train the classifier, we used the Euclidean distance between all pairs of strains as a one-dimensional feature and a binary encoding of within (1) or between (0) clade status as a model target. As there were far more pairs of strains from different clades, we measured classification accuracy with the Matthew's correlation coefficient (MCC), a metric that is robust to unbalanced counts in the confusion matrix (citation here). As a control, we compared the accuracy of each method's classifier to the MCC from a classifier fit to genetic distance between strains. t-SNE, UMAP, and PCA provided the most accurate classifications (MCC = 0.75, 0.67, 0.67, respectively) and outperformed pairwise genetic distance (MCC = 0.60) Figure 3 tbl. 1. MDS performed poorly (MCC = 0.43), confirming our expectations it would mirror genetic distances MCC value based on MDS's linear relationship with genetic distance. These results show the potential benefits of using t-SNE embeddings for cluster analysis over the computationally simpler genetic distance, despite the t-SNE's lack of global linear relationships between strains.

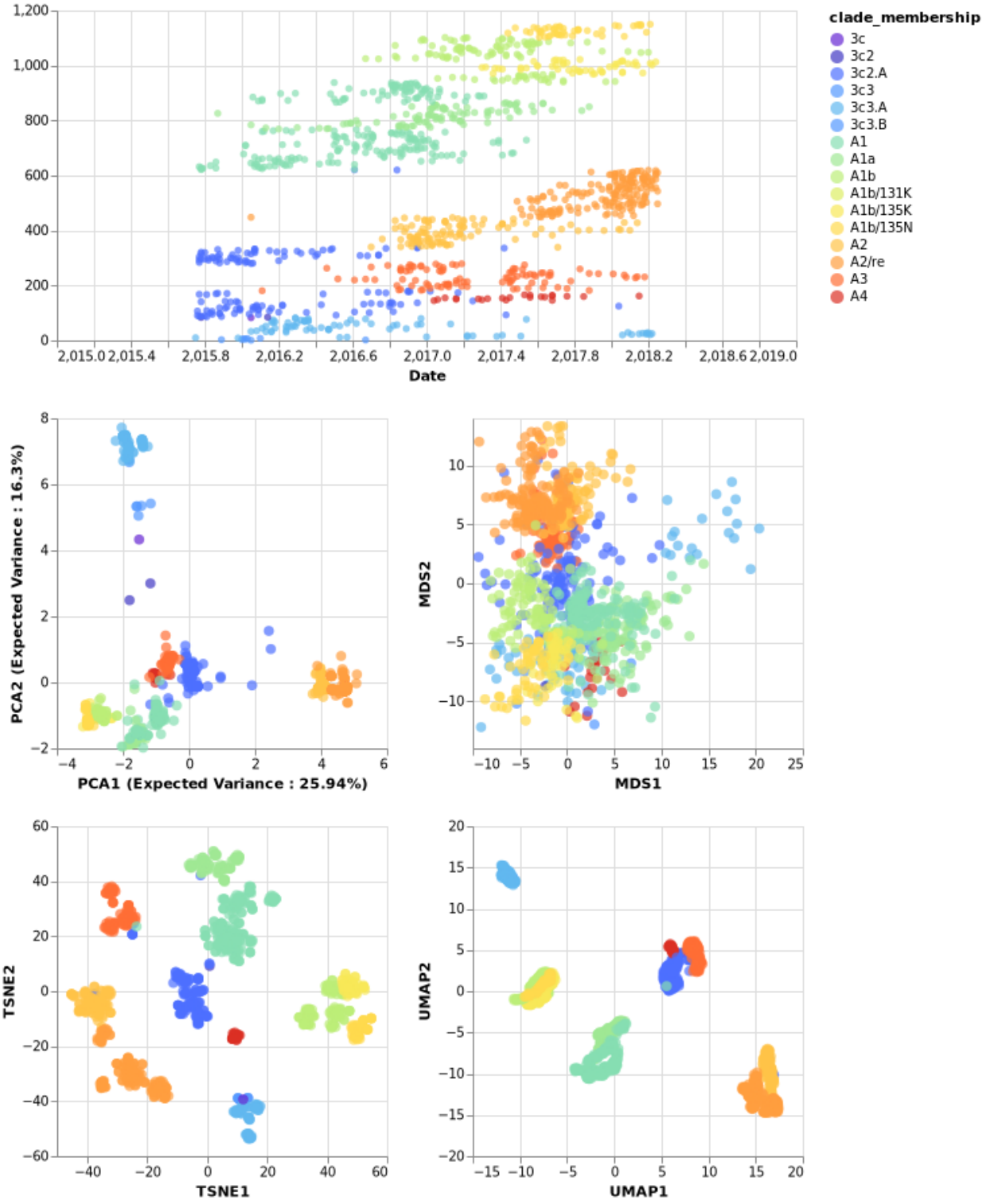


Figure 1: Genetic cartography of H3N2 strains by dimensionality reduction methods compared to inferred phylogeny.

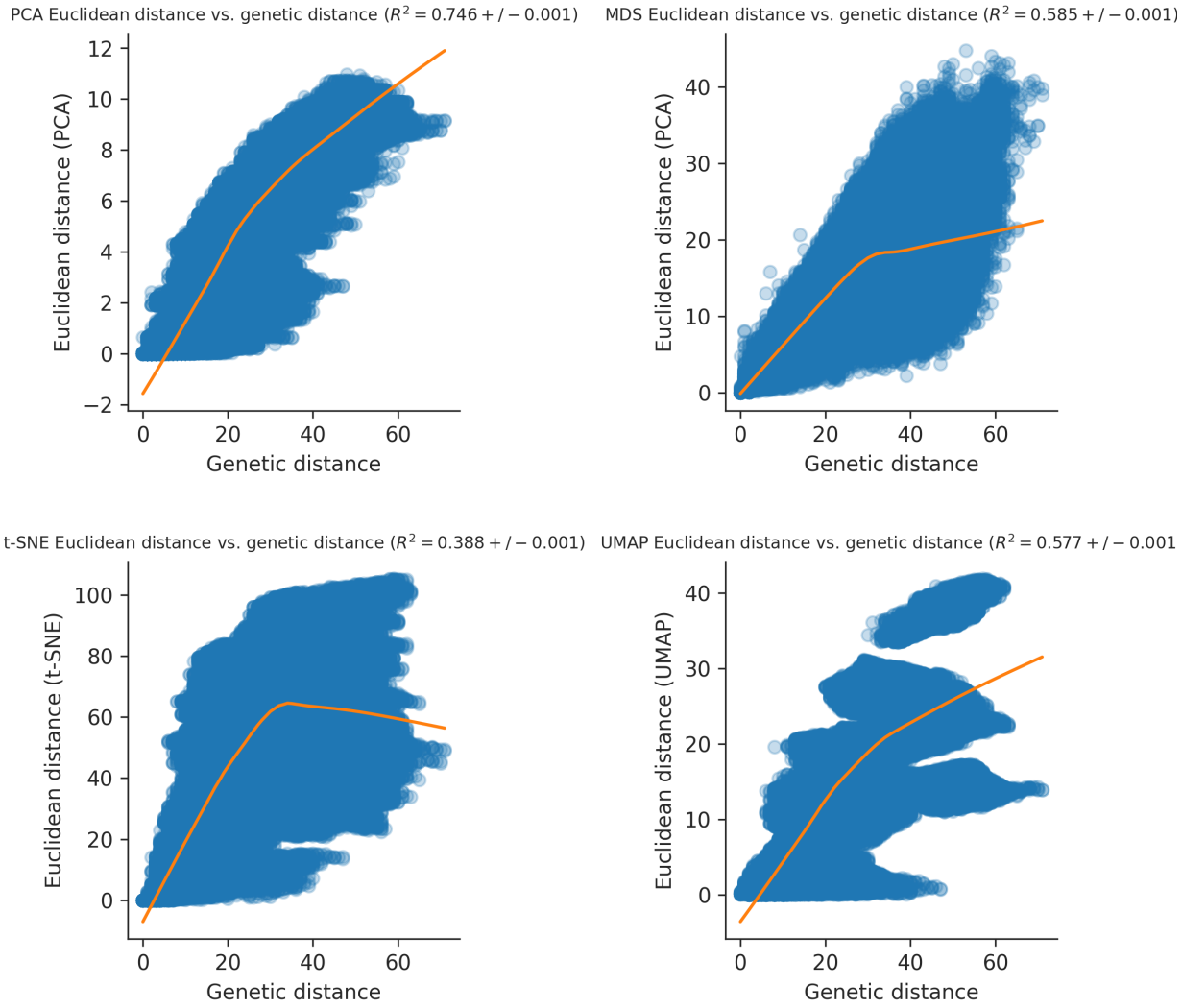


Figure 2: Mapping between Euclidean and genetic distances for all pairs of H3N2 strains by dimensionality reduction method.

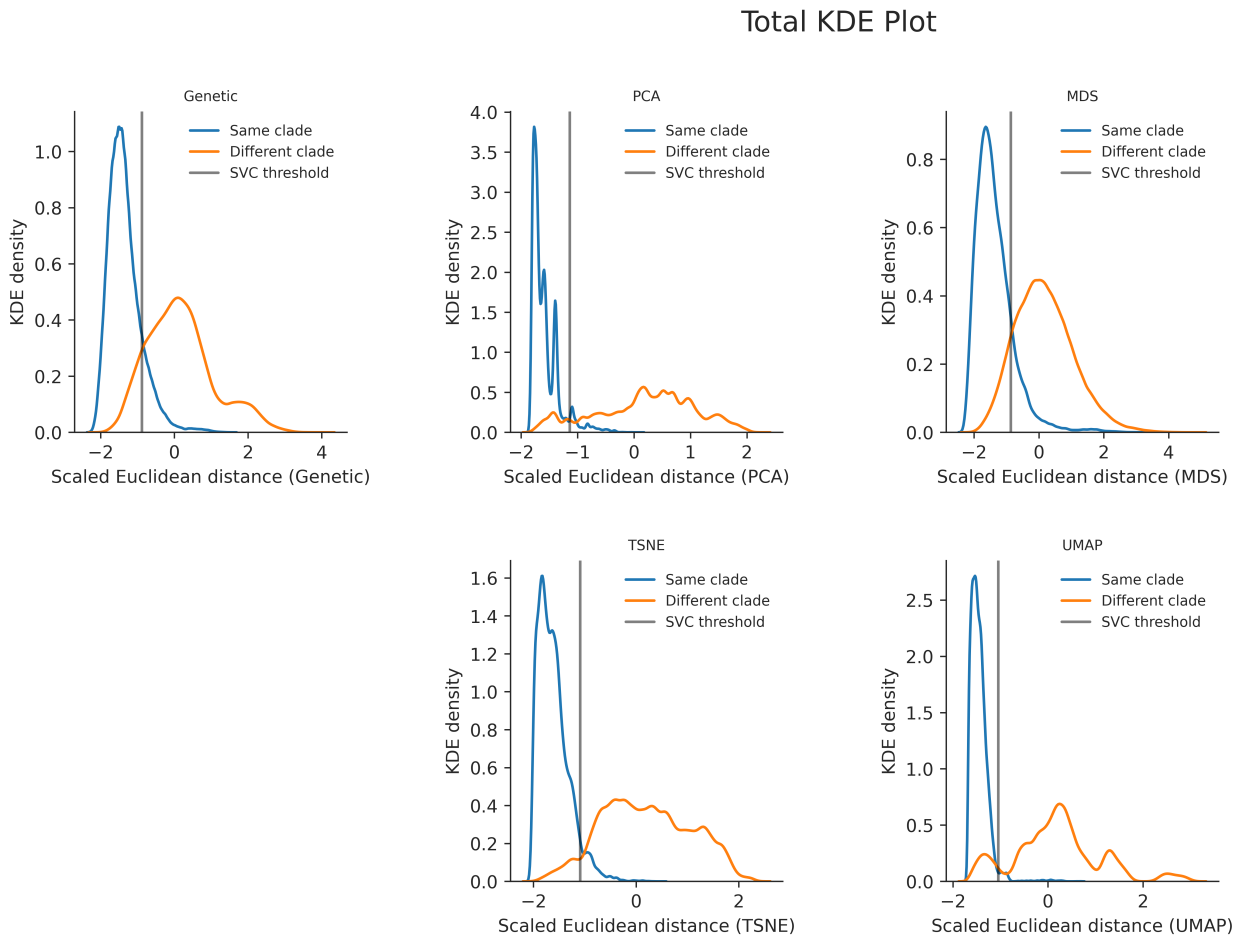


Figure 3: Distribution of scaled Euclidean distances between all pairs of H3N2 strains by clade status and dimensionality reduction method.

Embedding clusters reveal outbreak and geographical patterns within Zika

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

PCA and t-SNE exhibited a piecewise linear relationship for pairs of strains that differed by no more than 50 nucleotides (Figure 5). For larger than a 50 nucleotide difference in genetic distance, PCA, t-SNE, and UMAP's LOESS line becomes much steeper, revealing that these embeddings use local patterns to map genetically distant strain combinations farther away for better visualization. This is the expectation for t-SNE and UMAP, but is surprising to see in PCA. MDS provided the strongest linear mapping to genetic distance (Pearson's $R^2 = 0.738 \pm 0.001$). The UMAP mapping revealed two different clusters of points in the scatterplot, with the cluster at higher Euclidean distances dileneating the distance between clade c2 and the other strains due to its isolated sampling. This clustering is only seen in UMAP, revealing UMAP's sensitivity to outliers. t-SNE's mapping was fairly strong (Pearson's $R^2 = 0.52 \pm 0.002$) and revealed that pairs of strains that differ by more than 50 nucleotides are unlikely to be placed near each other in a t-SNE embedding.

Just as in influenza, t-SNE and PCA provided the most accurate classifications (MCC = 0.56 and 0.66, respectively) and outperformed pairwise genetic distance (MCC = 0.51) and UMAP (MCC = 0.37, Figure 6 tbl. 1). UMAP performed incredibly poorly, which we attribute to the incredible distance between clade c2 and the other clades, which may have caused the classifier to misrepresent the Euclidean threshold between and within clades (False Negative: 7934 vs False Positive: 49397 tbl. 1). MDS performed poorly (MCC = 0.41), confirming our expectation that MDS slightly underperforms genetic distance's classification, as MDS linerarly recapitulates genetic distance in euclidean space. These results corroborate our previous conclusion about the potential benefits of using t-SNE embeddings for cluster analysis over genetic distance.

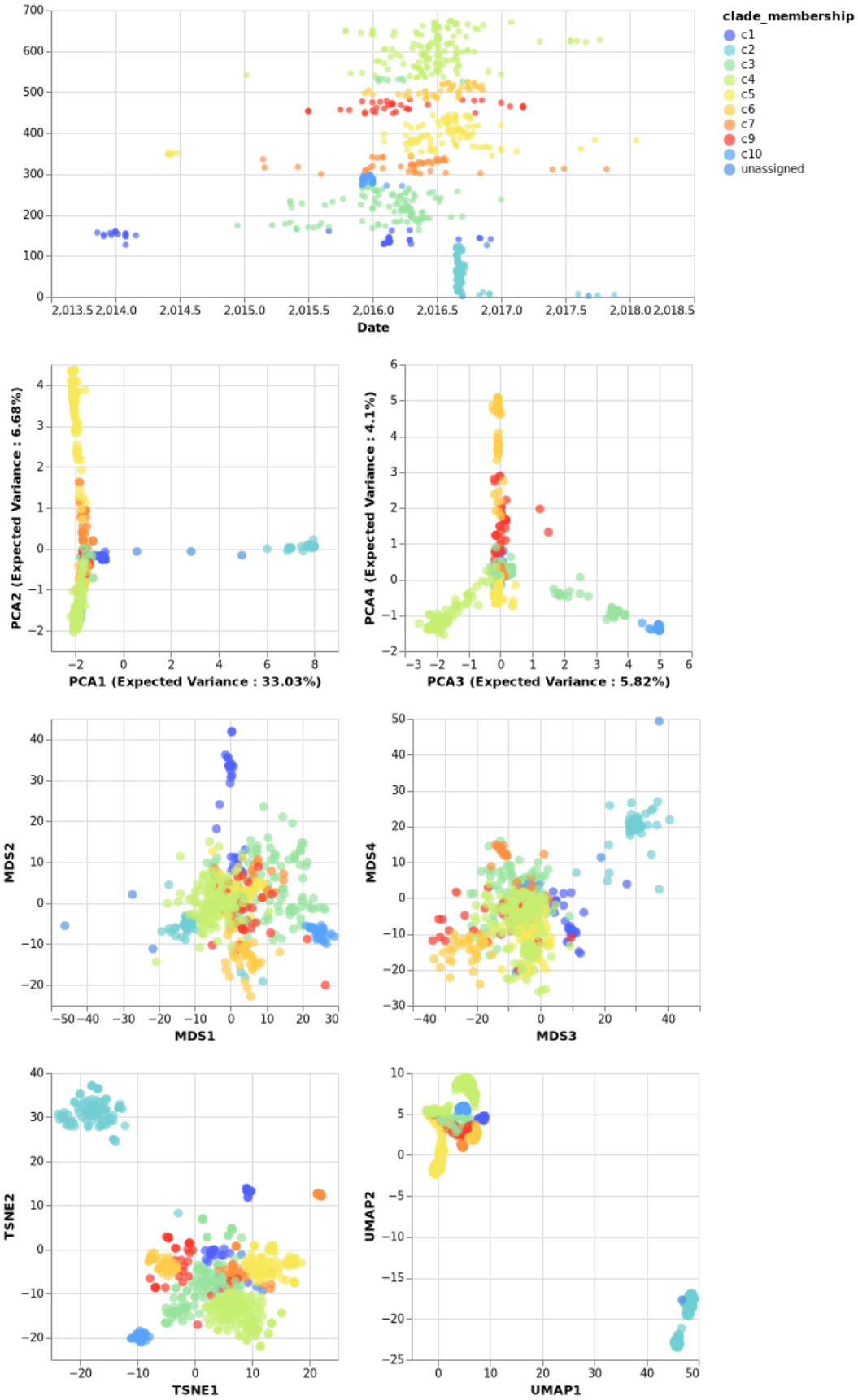


Figure 4: Genetic cartography of Zika strains by dimensionality reduction methods compared to inferred phylogeny.

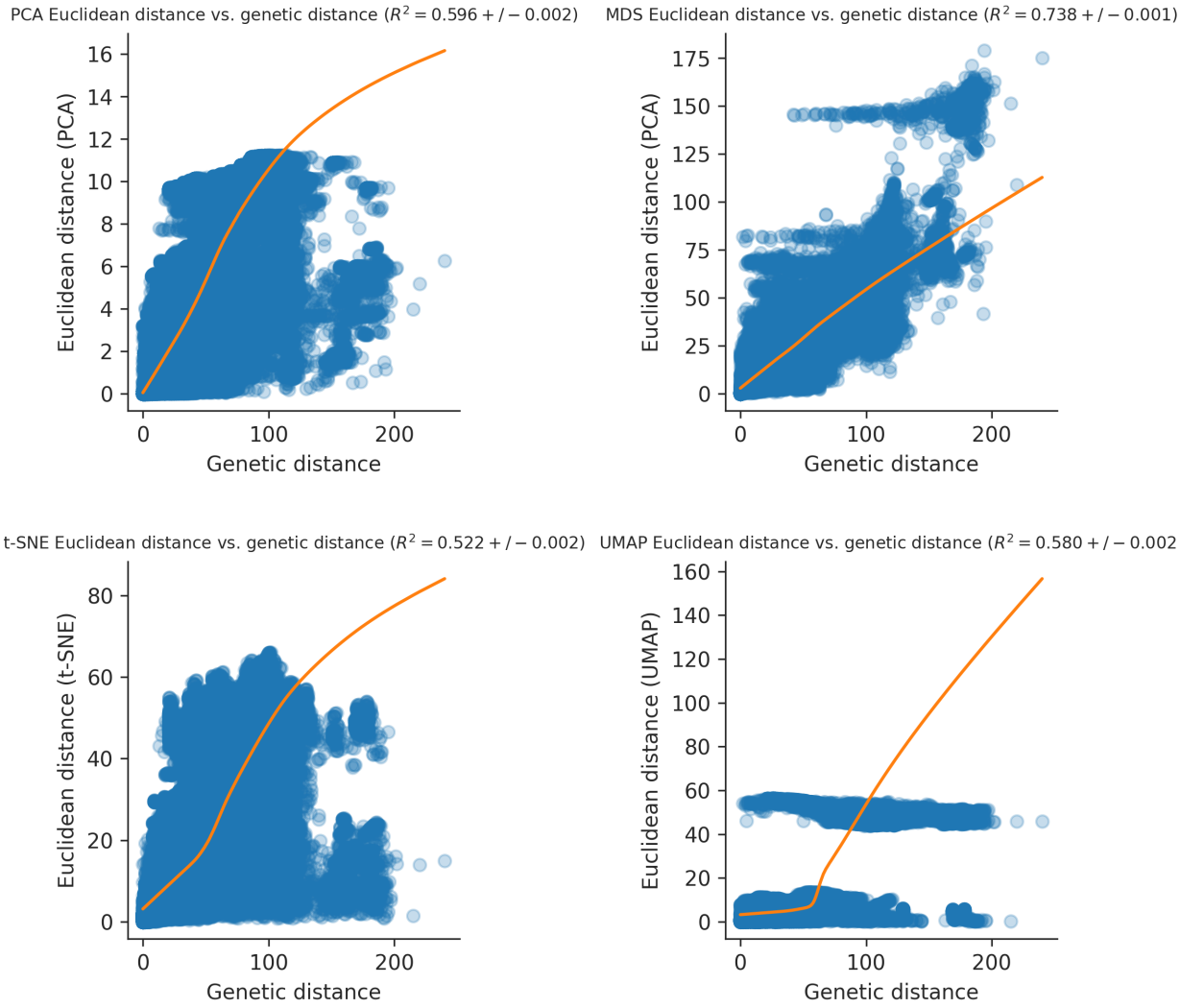


Figure 5: Mapping between Euclidean and genetic distances for all pairs of Zika strains by dimensionality reduction method.

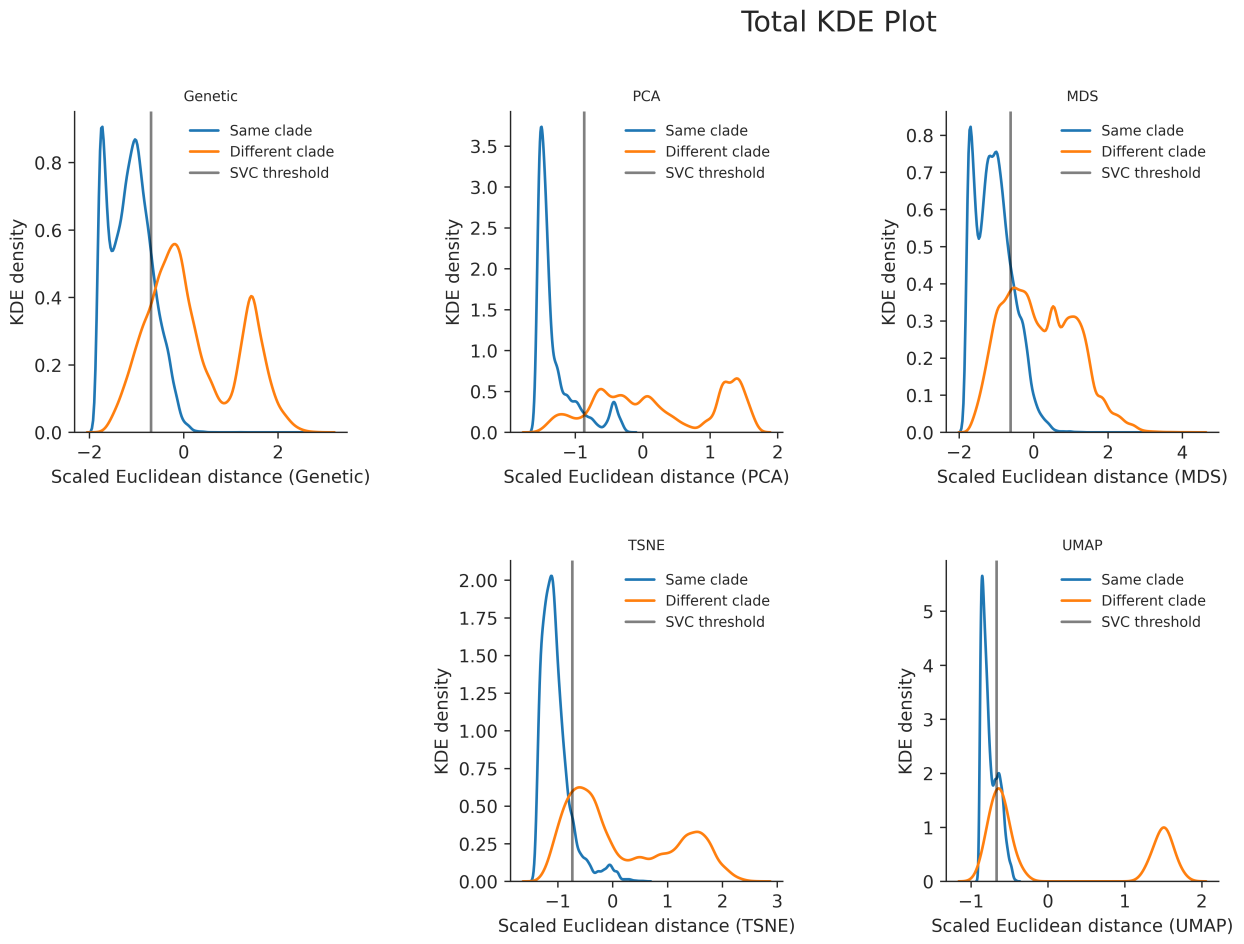


Figure 6: Distribution of scaled Euclidean distances between all pairs of Zika strains by clade status and dimensionality reduction method.

MERS within host outbreak patterns revealed with embedding clusters

While MDS, t-SNE, and UMAP recapitulated the patterns observed in the phylogeny, PCA did not (Figure 7). With MERS missing bases in multiple strains, imputation did not add more depth to the alignment. To combat this issue, all strains with missing bases 3 standard deviations higher than the mean were removed from the analysis; while this helped create tighter clusters in the distance based methods, PCA did not show any further patterns and its use was constrained to separating low quality strains from high quality. Isolated outbreaks and strains from different hosts were placed at larger Euclidean distances from related strains and hosts. Clade 32, a human outbreak from Seoul and surrounding territories isolated from other clades, is an outlier notably separated in t-SNE and UMAP and to a lesser degree in MDS. Because clades were defined by outbreaks that shared a common host, we expected local clustering within clades; this divergence was seen in t-SNE and UMAP in Clade 13, a human outbreak made up of two distinct clusters branching off from the same node in March of 2014. Clades 20, Clade 21, and Clade 22, camel outbreaks from Saudi Arabia, clustered together in all the distance based embeddings, reaffirming the genetic similarity of these strains. The embeddings clustered between hosts, with disease strains lacking a clade membership clustering into the nearest related clade. While t-SNE and UMAP's embeddings are very similar structurally, t-SNE performed exceedingly better at differentiating between intra-host clades highly related in the embedding. An example of this is the clear separation in t-SNE of the camel and human outbreaks concentrated in Saudi Arabia and the UAE (Clade 9, Clade 10, Clade 11, and Clade 12) Figure 8. This suggests that t-SNE is a strong tool for

viewing genetically homogeneous populations.

MDS had a linear relationship throughout while t-SNE and UMAP exhibited a piecewise linear relationship for pairs of strains that differed by no more than 100 nucleotides (Figure 9). For larger than 100 nucleotide differences, t-SNE and UMAP's LOESS lines decrease sharply, a contrast to the patterns seen in Zika and influenza. We see patterns of relatedness between UMAP and t-SNE's scatterplots in terms of shape, spread, and clustering (Pearson's $R^2 = 0.20 \pm 0.005$ and 0.25 ± 0.005 respectively). The two clusters in their scatterplots at similar genetic but different euclidean distances reveals that t-SNE and UMAP select a low or high Euclidean distance depending on the strain relationship (Clades 27 through 32 in one cluster, the rest in the other). MDS provided the strongest linear mapping to genetic distance (Pearson's $R^2 = 0.76 \pm 0.003$). The same mapping for PCA was incredibly weak and nonlinear (Pearson's $R^2 = 0.023 \pm 0.001$).

Just as in influenza and Zika, t-SNE provided the most accurate classification of the embeddings (MCC = 0.63), but did not outperform pairwise genetic distance (MCC = 0.71, Figure 10 tbl. 1). PCA could not be considered in this analysis, as the classifier was unable to find a distance threshold to separate within vs between clade relationships. UMAP performed poorest (MCC = 0.48), which we attribute to the tight clusters between multiple related clades seen in UMAP, which may have caused the classifier to create a lower Euclidean distance threshold between and within clades (False Negative: 309 vs False Positive: 2164) tbl. 1. MDS performed much better than in influenza and Zika (MCC = 0.62) and minorly underperformed genetic distance, which we conclude is because of the proportional relationship between genetic and euclidean distance as seen in the scatterplot. Because all MDS distance calculations are computed for its four leading components, more variance is explained in the supplemental plots than the interactive chart (which explains why MDS is equivalent to t-SNE for classification within MERS). These results corroborate our conclusion about using t-SNE embeddings for cluster analysis, but suggests viewing and quantifying the data through multiple reductions in order to create the best view of the data.

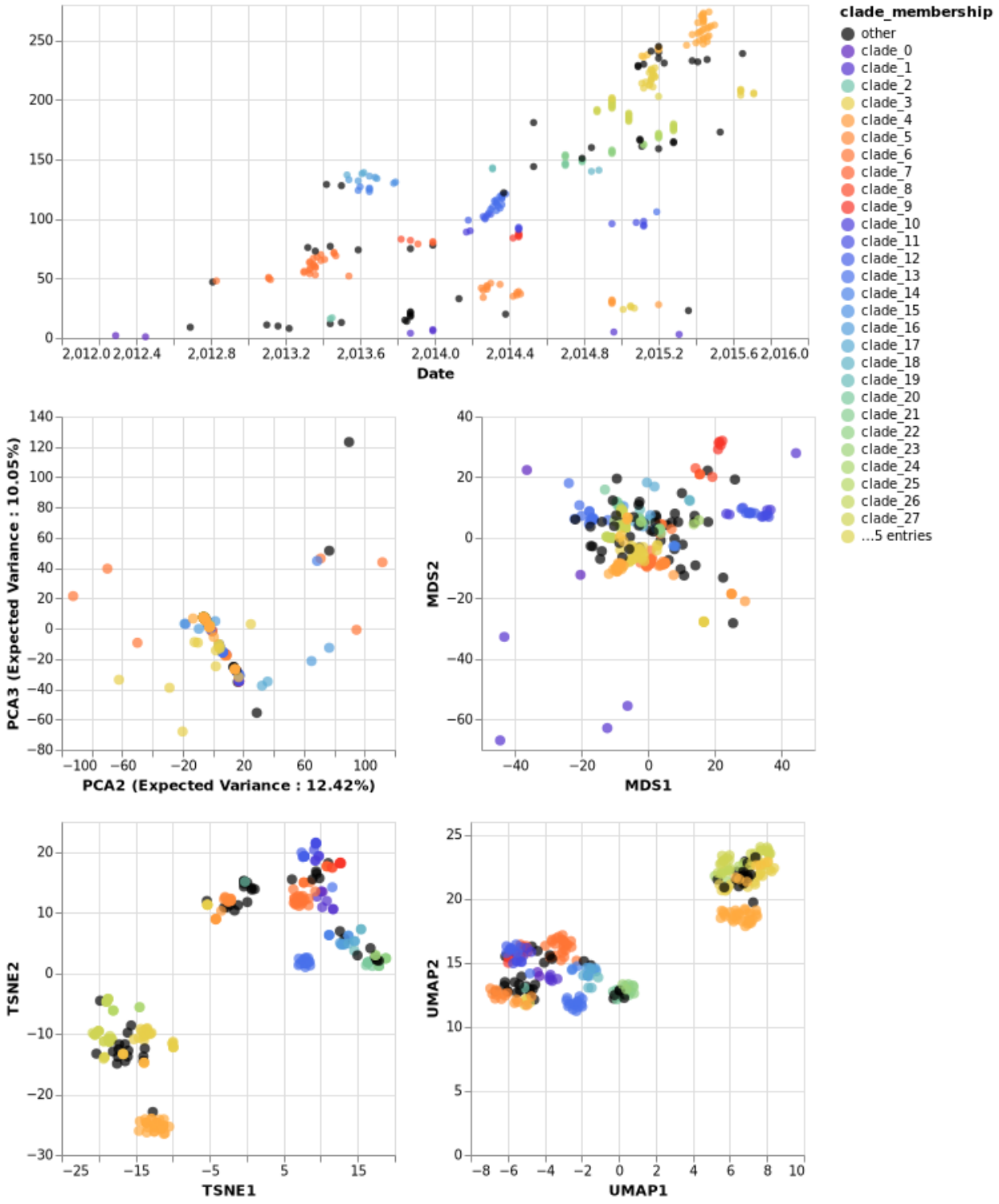


Figure 7: Genetic cartography of MERS strains by dimensionality reduction methods compared to inferred phylogeny.

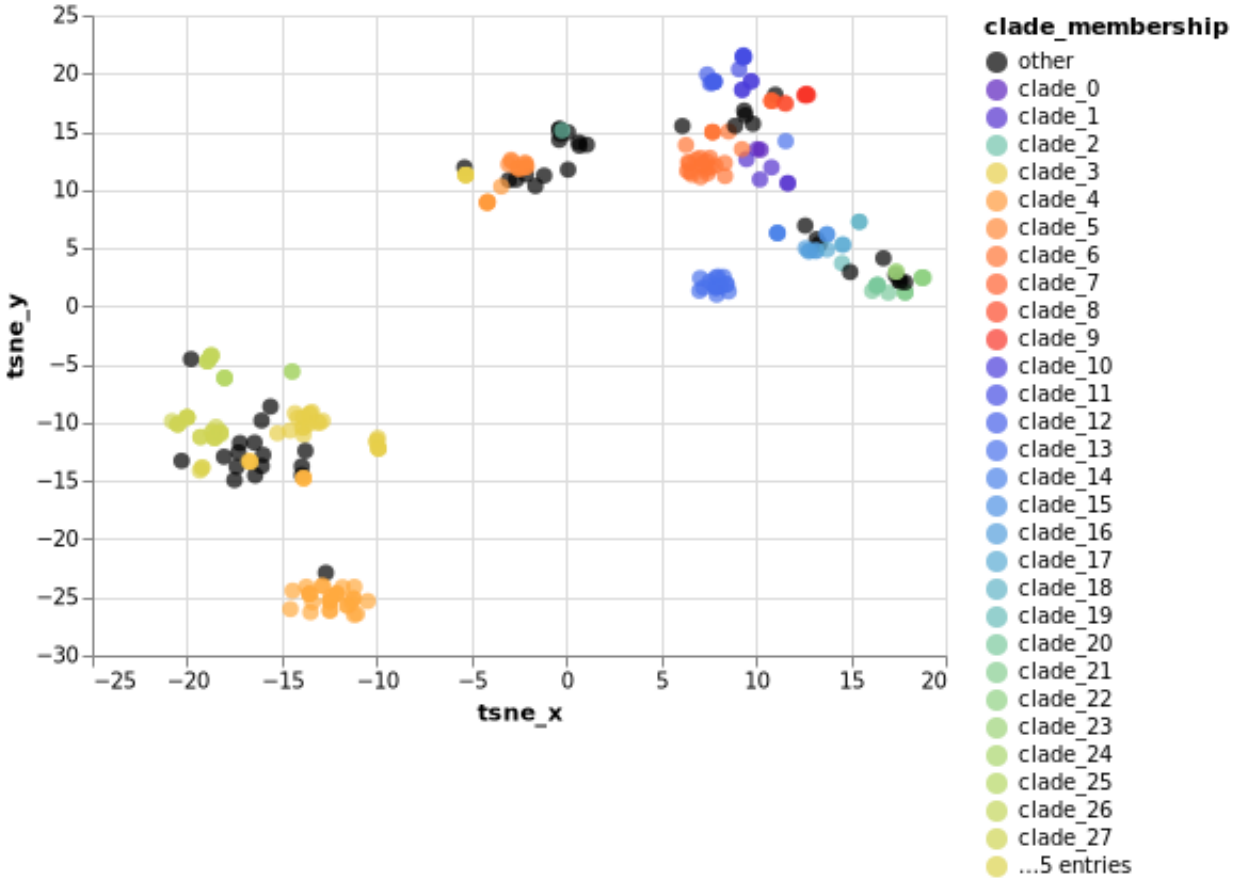


Figure 8: Interactive t-SNE chart; zoomable with interactive tooltips

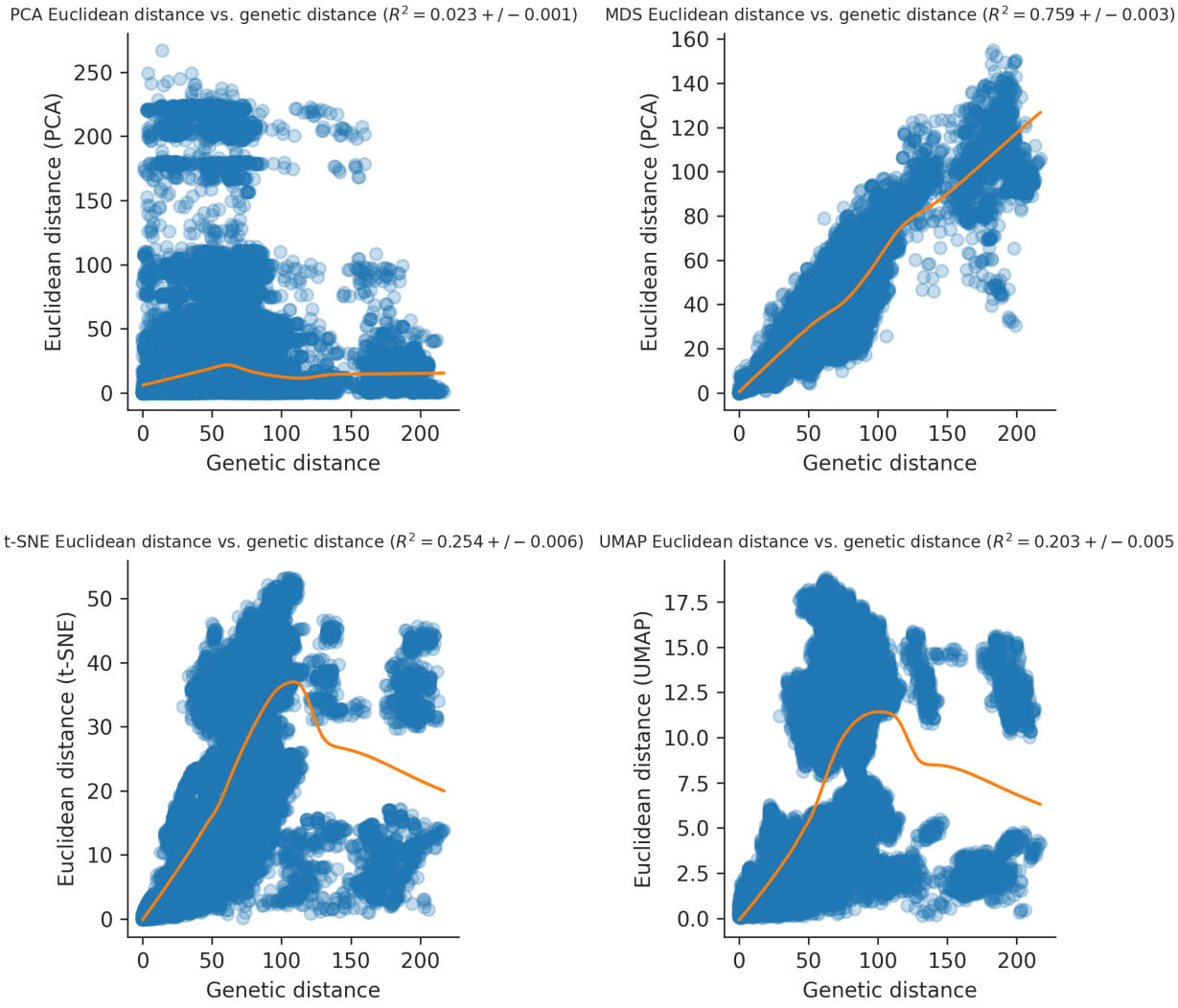


Figure 9: Mapping between Euclidean and genetic distances for all pairs of MERS strains by dimensionality reduction method.

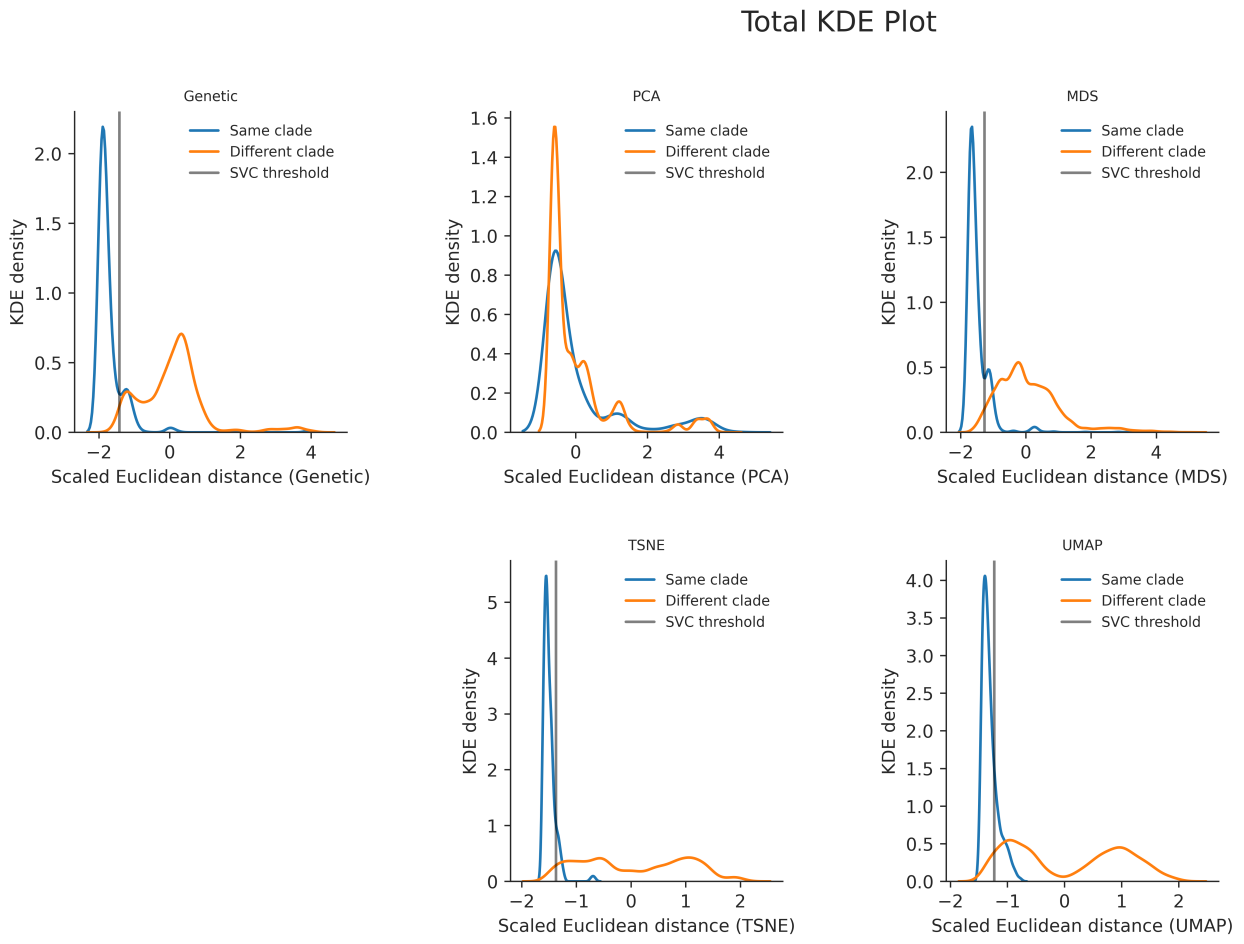


Figure 10: Distribution of scaled Euclidean distances between all pairs of MERS strains by clade status and dimensionality reduction method.

Discussion

In this paper, we analyzed the usage of PCA, MDS, t-SNE, and UMAP to better understand population structure in varying types of diseases with differing clade assignments. We accomplished this by using interactive visualizations, a novel and important part of the analysis that enables a more in-depth exploration of patterns in the data that would otherwise require more technical expertise. This interactibility makes these charts, along with the public automated code pipelines and scripts from this paper, more accessible to scientists and the public.

Across all the diseases, we found PCA to be incredibly sensitive to missing data. While imputing missing data and dropping strains creates a more correlated PCA embedding, it also introduces noise. We attributed this drawback to the larger flaw of using sites on a genome as features to find useful patterns. This, however, reveals that PCA is useful for finding low quality and misnamed strains in a population. An advantage of distance based metrics over PCA is their robustness to missing data while preserving similar, if not better, quality results.

A question we aimed to answer with this research was the level of the phylogeny at which clusters were developing within each embedding. Because t-SNE performed the best at finding outbreaks, newly created clades, and local patterns, we used the virus populations' phylogenies colored by t-sne x (component 1)

to analyze Flu, Zika, and MERS clade structure and coloring. Across all three viruses, t-SNE pulled out outbreaks as fine as groups of 4 strains or less, which were at most one node away from each other. This level of fine-tuning to groups defined by geographical introductions revealed t-SNE's ability to reference ancestral population structure in structuring the embedding. Of the four embeddings, the best indicator for quantitative determination of clade status (how many relationships are preserved) was t-SNE's Euclidean distance, while genetic distance's MCC was higher than t-SNE's for MERS. As genetic variance and genome length continues to increase, t-SNE continues to work well, even with added noise creating less segmented embeddings.

There are limitations to this research. The inherent flaw of these embeddings is their failure to account for likelihood of ancestry and infer ancestral relation. While this makes the embeddings much quicker for identifying genetic relationships within populations, not having ancestral relationship information means this approach should be used in conjunction with phylogenetic trees (provide quality control), and in rapidly emerging viruses, where these embeddings would provide information about where sequences cluster quickly and accurately. Computationally, there are limitations to the embedding algorithms themselves. The algorithms used to reduce the data are unsupervised learning techniques. While they are powerful tools for viewing larger scale patterns within a population, the inherent noise in these renderings of multidimensional data means the data does not always cluster in biologically or statistically meaningful groups. There is also inherent bias and skew from missing and incomplete data, which adds noise into these algorithms. A future direction for this research would be using a semi-supervised learning method. In this approach, the embedding obeys a certain ground truth, where a subset of expert-reviewed strains are labeled, and that data trains the parameters of the embedding. The rest of the unlabeled data would then be projected into this new environment, which has been proven to greatly improve both the accuracy and usefulness of the visualizations. The limitation to this approach is the time and computational work required for researchers to create an expert approved dataset diverse enough for a strong semi-supervised learning approach.

There are many other directions to further test the scope and applicability of these embeddings. One such direction is a cross-validation test, where a threshold of Euclidean distance between relationships is trained on a population from the past, and tested on a population of the same disease from some years later. This could potentially determine the embedding's abilities to understand future trends in present populations. Another future direction for this research would be finding parameters for the embeddings that work best across viruses. This research found parameters within influenza that we inferred would work across the rest of the analysis for consistency. If these parameters, once calibrated for the best results across all the viruses consistently returned similar values, could mean there is a set of parameters that works best for evolutionary data. This calibration would be useful when defining a Euclidean threshold between outliers and normal strains in order to quickly find outliers within a population. As it is common practice to run flu builds all the way through, view them, and find a handful of outliers that should now be excluded from the analysis, this research could potentially be used as an upstream tool to flag potential outliers before building a tree to potentially save scientists hours of work in outlier detection.

Our recommendations for algorithm choices are influenced by the many factors that make up a population. For a population with smaller genomes and very few missing bases, PCA and t-SNE work best for defining clades and conveying useful information. For genomes with around 10K bases and more missing data, PCA still works with basic imputation; the caveat, however, is the added noise. To circumvent this issue, t-SNE creates a similar, better quality embedding than PCA without imputation, making it a more versatile algorithm. Classification of within and between clade relationships is best performed by t-SNE or genetic distance. For large genomes (30K bases+), lots of missing bases, fewer samples, and diversity, t-SNE is the most useful for viewing qualitative patterns. Genetic distance works best in these populations for classifying relationships as within or between clade. A visualization of the raw pairwise data would best be achieved through MDS, as MDS consistently creates the strongest correlated linear relationships between pairwise and Euclidean distance. While UMAP has been used extensively recently in genomic studies, we recommend the use of t-SNE for an embedding more robust to outliers and other sample quality issues that are extrapolated within the UMAP embedding.

This paper has systematically and quantitatively demonstrated the usefulness, accuracy, and usages of these embeddings in viral epidemiology, something not done until now. It has delved deeper into the scope of these

embeddings by analyzing the results from well known viruses, and created tools used in this paper that are public and easy to use on other datasets. We hope scientists will now be able to use these embeddings with confidence to further understand their dataset and viruses, and use them to quickly and accurately view strain relationships within rapidly emerging diseases such as SARS-CoV-2, where a lack of data and samples can make traditional epidemiological tools less useful.

Materials and Methods

The analysis environment can be recreated using conda and all installation instructions are available on this paper's github .

The genome data we used for H3N2 HA influenza is from the NCBI influenza database. We used this search. Clades were defined by reasonable phylogenetic signal. The Zika data was curated by Allison Black, with sequences from Genbank and the Bedford Lab. Clades were defined by regionally important introductions and reasonable phylogenetic signal. The MERS data was downloaded from e-life. (Dudas et al. 2018)

Clades and host were used in the MERS analysis, as the hosts (camel and human) are scientifically useful and phylogenetically accurate to the Newick tree. The clade assignments were defined based on monophyletic host status (strictly camel or human) to reveal patterns within host outbreaks. We analyzed influenza A/H3N2 and Zika by creating a FASTA file of multiple sequence alignments with MAFFT v7.407 (Katoh et al. 2002) via augur align (Hadfield et al. 2018) and phylogenies with IQ-TREE v1.6.10 (Nguyen et al. 2014) via augur tree version 9.0.0.

We used two different methods of transforming the data; Scaling and centering the data, and a Hamming distance similarity matrix. For Scaling and Centering the data, we performed PCA on the matrix of nucleotides from the multiple sequence alignment using scikit-learn (Jolliffe and Cadima 2016). An explained variance plot was created to determine the amount of PCs used for distance calulations and visualization, which is in the supplementary figures section. A separate bases missing vs PC1 was also created to help reveal the level of relation between missing bases and outliers in PCA; this is available for MERS in the supplemental section.

We dropped around 4 strains in the H3N2 analysis, as they were direct animal to human transmissions where the genomes resembled swine flu (seen through NCBI's BLAST). We dropped around 5 strains in the Zika analysis that were exceedingly low quality. Due to the amount of missing data within the zika genome, we also imputed the data using scikit-learn's simple imputer for PCA for a better embedding result. This was only applied to PCA, as the hamming distance algorithm used with the distance based methods disregards missing bases. Imputation was tested for MERS, but due to entire columns of missing data for MERS, we dropped all strains with over 3 standard deviations of missing bases in its genome from the MERS analysis.

For Hamming distance, we created a similarity matrix. By comparing every genome with every other genome and clustering based on their Hamming distance, distance-based methods take the overall structure of the multidimensional data and groups together genomes with similar differences. This means the data is clustered by genetic diversity (in a phylogenetic tree genetic diversity is categorized using clades). Each genome was split into separate nucleotides and compared with other nucleotides in the same site on other genomes. We only counted a difference between the main nucleotide pairs (AGCT) - gaps (N, -, etc.) were not. This is because some sequences were significantly shorter than others, and a shorter strain does not necessarily correspond to genetic dissimilarity, which is what counting gaps implied.

We reduced the similarity distance matrix through MDS, t-SNE, and UMAP, plotted using Altair (VanderPlas et al. 2018), and colored by clade assignment. Clade membership metadata was provided by a .json build of the influenza H3N2 tree and Zika trees. For MERS, the host data was given via the Newick tree, and clade membership was defined using BioPython as outbreaks with a monophyletic host status (strictly camel or human).

The 3 different dimensionality reduction techniques are ordered below by publication date: - MDS - t-SNE - UMAP

The plots of the full 10 PCs for PCA and the first 6 components for MDS are available in the supplemental figures section.

We tuned hyperparameters for t-SNE and UMAP using an exhaustive grid search, which picked the best parameters by maximizing Matthews Correlation Coefficient for the confusion matrix created from the Supported Vector Machine's classification. UMAP's minimum distance and nearest neighbors were tuned, and t-SNEs perplexity and learning rate were tuned. As nearest neighbors fluctuates depending on the amount of samples, we took the best nearest neighbor value from the cross validation and the total number of samples given per fold. This proportion was used to determine the nearest neighbors value for the UMAP plots. t-SNE performed best with a perplexity of 15.0 and a learning rate of 100.0. UMAP performed best with a minimum distance of .05 between clusters. While tuning these parameters does not change qualitative results, it can help make patterns easier to identify.

We ran the raw embedding distances through the clustering algorithm Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to understand the usage of the embeddings to cluster data without the phylogenetic tree.

To further analyze these embeddings' ability to accurately capture the multidimensional data, we made two separate plots: Hamming vs Euclidean distance scatterplots with a LOESS best fit line, and within vs between clade KDE density plots per embedding.

Hamming distance vs Euclidean distance scatterplots: Hamming distance vs Euclidean distance plots assess the local and global structure of the embedding as well as assess the overall strength of the embedding recapitulation. The Hamming distance between nucleotide sequences is plotted on the x axis, and the Euclidean distance between the points in the embedding are plotted on the y axis. PCA and MDS's distances were calculated using 4 components, while t-SNE and UMAP were calculated with 2. By plotting these distance measurements, we can observe how correlated the dataset is. The higher the correlation, the better a function can describe the relationship between the Hamming distance value and the Euclidean distance value. In this way, constant correlation in a plot reveals that the embedding tends to capture and retain global patterns, and a splayed structure points to local structure preservation. Therefore, the closer the Pearson Coefficient is to 1, the better the embedding is at preserving pairwise relationships in Euclidean space. The LOESS line drawn through the plot assesses the best fit function for the embedding. We bootstrapped our scatterplot to find the Pearson Coefficient with a confidence interval.

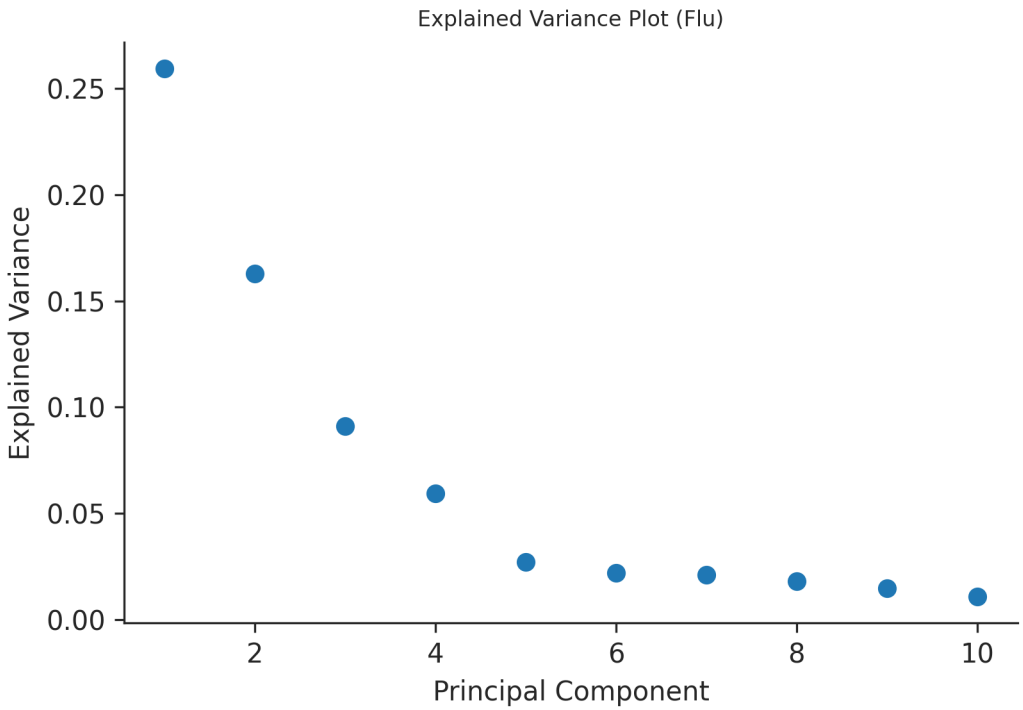
Between vs Within clade KDE Density Plots: The Between vs Within clade KDE Density Plots visually represent how well Euclidean distances can distinguish virus genomes from different clades. In other words, it describes the probability that a certain Euclidean distance can be used to classify a given pair of genomes as within vs between clades. The larger the median ratio between the two curves presented per clade relationship, the higher the relative probability that the embedding will accurately predict if two strains with any specific distance is a between or within clade relationship. To create this plot, the matrix of Euclidean distances for each embedding was flattened, and each comparison was labeled as a "within clade" or "between clade" comparison using the clade assignments from the .json build of the tree. KDE plots were made using seaborn, separated by clade status and Euclidean distance on the y axis. A Supported Vector Machine was run to optimize for clade relationships by Euclidean distance, and the Matthews Correlation Coefficient, accuracy value, and classifier thresholds were calculated and captured along with the confusion matrix of values.

Acknowledgements

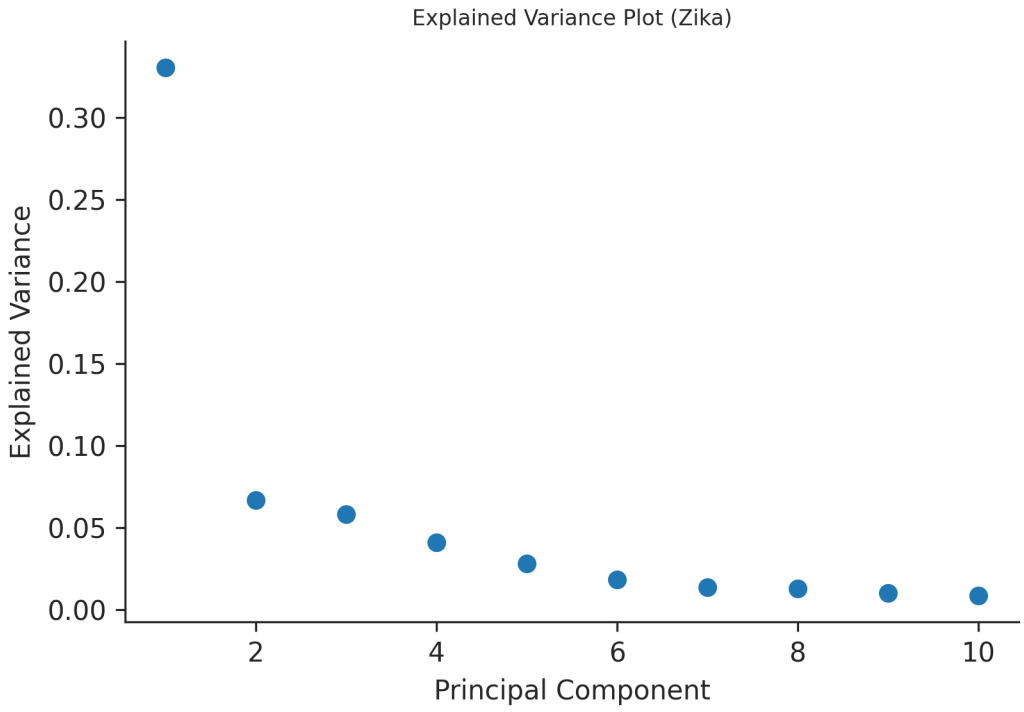
I thank the Bedford Lab at the Fred Hutch, specifically Dr. Trevor Bedford, for the opportunity to research and publish work for this project. For help received through the lab, I thank John Huddleston, for his code reviews, edits, weekly meetings, observations, and time to help make me a better scientist. I thank Allison Black for curating the samples and phylogeny for Zika.

Supplementary Figures and Analysis

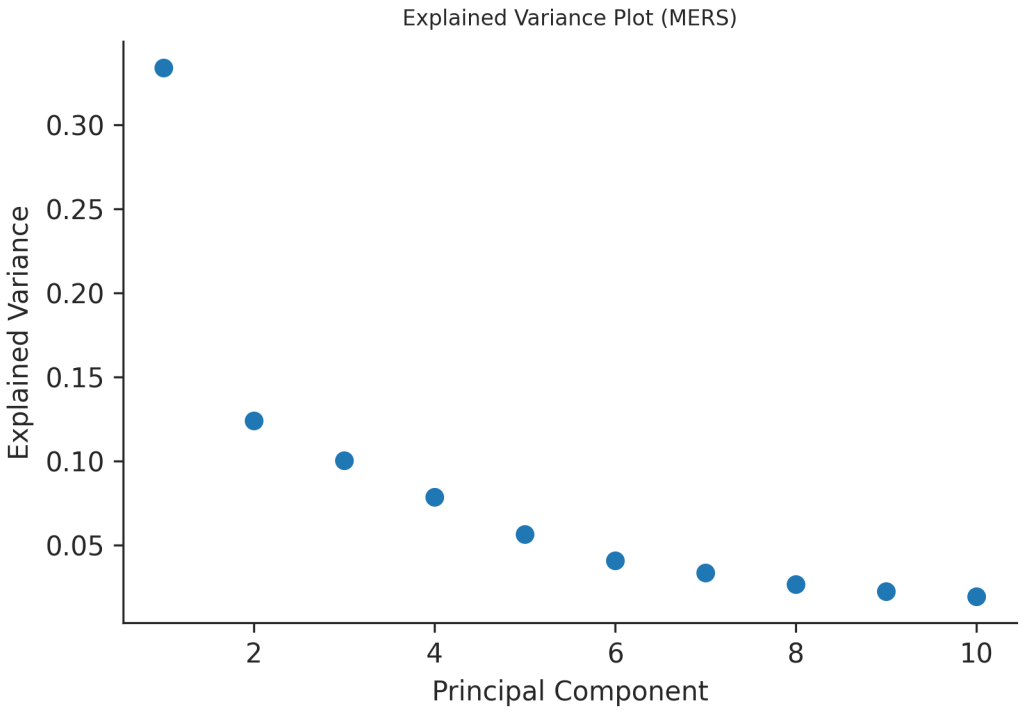
Explained Variance Plots for PCA



Flu



Zika



MERS

PCA Full Plots

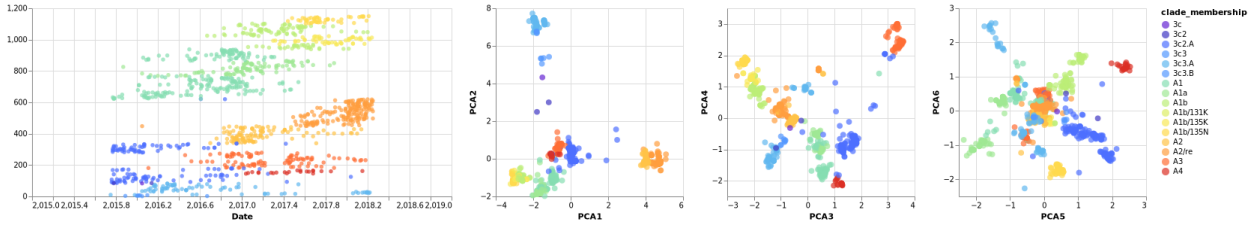


Figure 11: PCA Full Plot - Flu

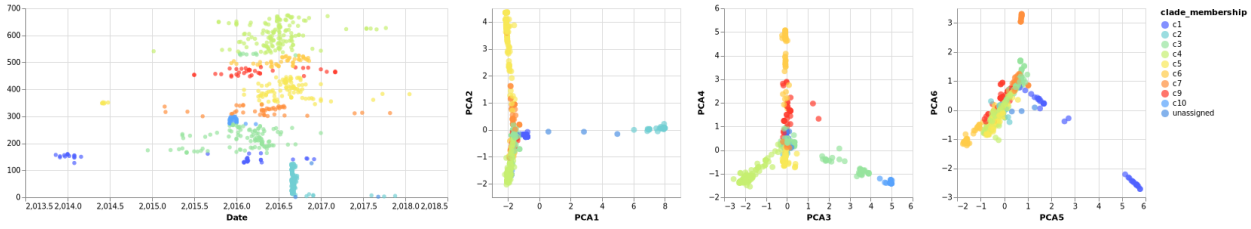


Figure 12: PCA Full Plot - Zika

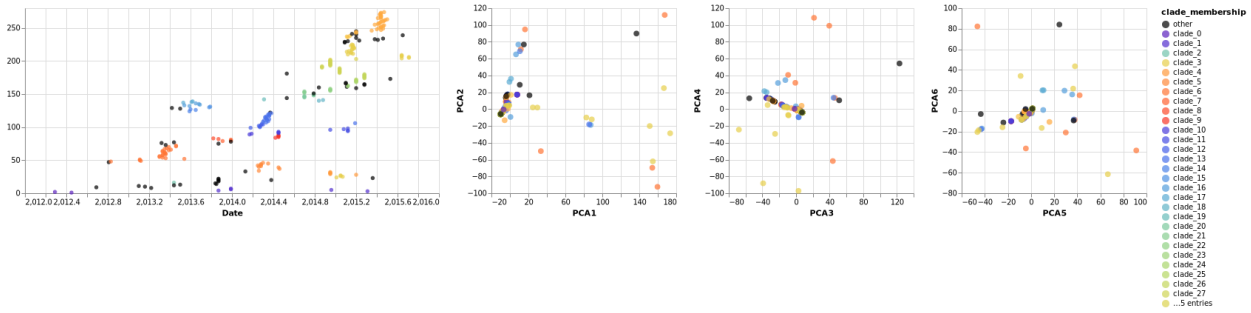


Figure 13: PCA Full Plot - MERS

MDS Full Plot:

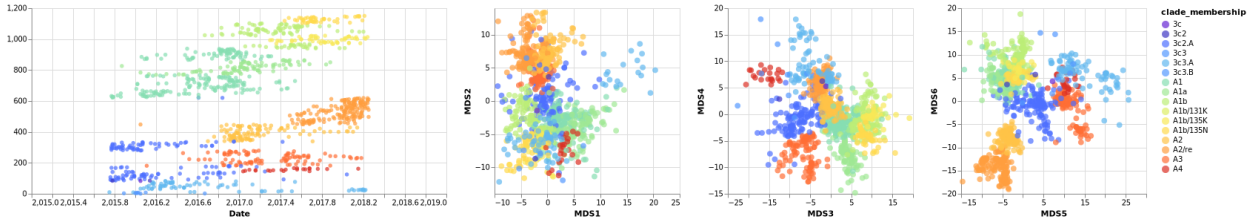


Figure 14: MDS Full Plot - Flu

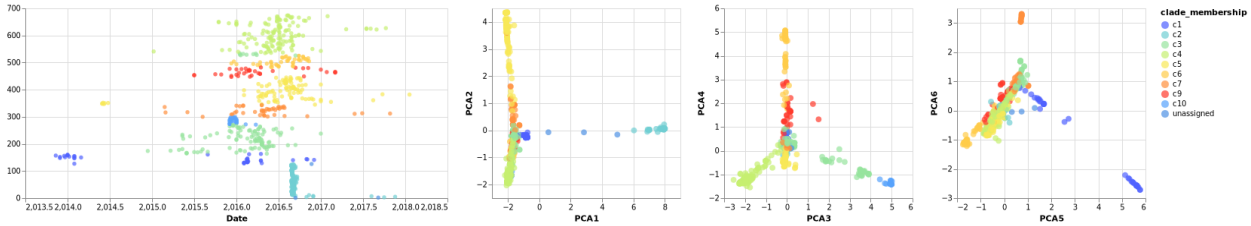


Figure 15: MDS Full Plot - Zika

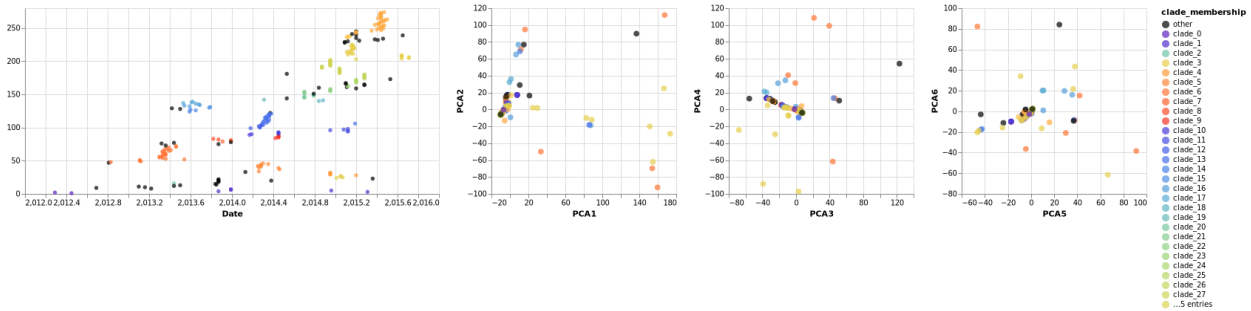
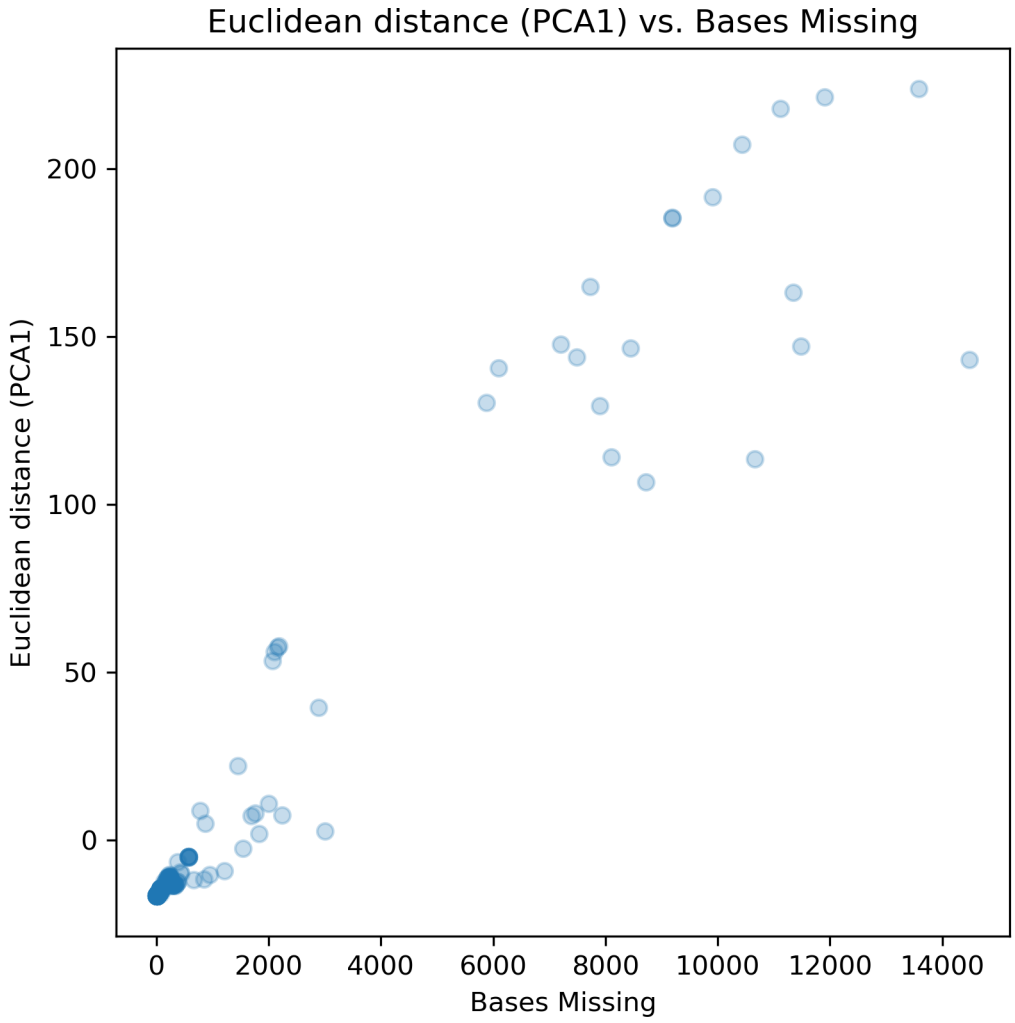


Figure 16: MDS Full Plot - MERS

Bases Missing VS PC1 Plot:MERS



Works Cited

Alexander, David H, John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research*.

Diaz-Papkovich, Alex, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel. 2019. "UMAP Reveals Cryptic Population Structure and Phenotype Heterogeneity in Large Genomic Cohorts." *PLOS Genetics*, November. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008432.

Dudas, Gytis, Luiz Max Carvalho, Andrew Rambaut, and Trevor Bedford. 2018. "MERS-CoV Spillover at the Camel-Human Interface." *eLife*, January. https://elifesciences.org/articles/31257.

H. C., Metsky, Matranga C. B., Wohl S., Schaffner S. F., Freije C. A., Winnicki S. M., West K., et al. 2017. "Genome Sequencing Reveals Zika Virus Diversity and Spread in the Americas." *Nature*. https://doi.org/10.1038/nature22402.

Hadfield, J., C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. 2018. "Nextstrain: Real-Time Tracking of Pathogen Evolution." *Bioinformatics*, May, bty407. https://doi.org/10.1093/bioinformatics/bty407.

Hout, Michael C., Megan H. Papesh, and Stephen D. Goldinger. 2012. "Multidimensional Scaling." *Wiley Online Library*.

Jolliffe, Ian T, and Jorge Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." *Nucleic Acids Research* 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.

Kosakovsky Pond, Sergei L, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon D W Frost. 2006. "Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm." *Molecular Biology and Evolution*.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9 (Nov): 2579–2605.

Martin, Darren P, Ben Murrell, Arjun Khoosal, and Brejnev Muhire. 2017. "Detecting and Analyzing Genetic Recombination Using Rdp4." *Methods in Molecular Biology (Clifton, N.J.)*.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." http://arxiv.org/abs/1802.03426.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2014. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, et al. 2008. "Genes Mirror Geography Within Europe." *Nature*.

Peter H. Sudmant, Eugene J. Gardner, Tobias Rausch. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature*, September.

Pérez-Losada, Marcos, Miguel Arenas, Juan Carlos Galán, Ferran Palero, and Fernando González-Candelas. 2015. "Recombination in Viruses: Mechanisms, Methods of Study, and Evolutionary Consequences." *Infection, Genetics and Evolution*.

Posada, David, and Keith A. Crandall. 2001. "Evaluation of Methods for Detecting Recombination from DNA Sequences: Computer Simulations." *Proceedings of the National Academy of Sciences* 98 (24): 13757–62. https://doi.org/10.1073/pnas.241370698.

Rambaut, Andrew, Oliver G. Pybus, Martha I. Nelson, Cecile Viboud, Jeffery K. Taubenberger, and Edward C. Holmes. 2008. "The Genomic and Epidemiological Dynamics of Human Influenza a Virus." *Nature*, April. https://www.nature.com/articles/nature06945.

VanderPlas, Jacob, Brian E. Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. "Altair: Interactive Statistical Visualizations for Python." *Journal of Open Source Software* 3 (32): 1057. https://doi.org/10.21105/joss.01057.