Genetic cartography reveals ancestral relationships of human pathogenic viruses

true true true true

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 Cartography's 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 invalidates basic phylogenetic assumptions of inheritance and requires more sophisticated approaches. 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, Novembre, and Lange 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 as to 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 visualization of the embeddings for a deeper understanding and exploration of patterns within and 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.

Recombination: occurs when two or more viral genomes co-infect the same host cell and exchange genetic segments. Shuffling/reassortment: a type of recombination in viruses with segmented genomes where complete genome segments are interchanged to increase genetic variety through the creation of new segment combinations (Pérez-Losada et al. 2015).

Results

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

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 cannot be used on a similarity matrix, and is intended only for transformed and normalized multidimensional data. Multidimensional Scaling (MDS) refers to statistical techniques that increase the interpretability of local relational structures mired in the dataset (Hout, Papesh, and Goldinger 2012). A limitation to MDS is that only one symmetric matrix is allowed as input and the scale of measurement is non-numerical. MDS algorithm places higher importance on translating dissimilarity to distance than exacerbating patterns locally. t-distributed Stochastic Neighbor Embedding (t-SNE) largely focuses on local over global structure, and t-SNE's projection of clusters and distances between clusters 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 based in Riemannian geometry and algebraic topology (McInnes, Healy, and Melville 2018). The end result is low dimensional neighborhoods that group genetically similar strains together on a local scale while still preserving long-range connections between more distantly related strains. A limitation includes its lack of maturity - this novel technique does not have firmly established practices and robust libraries to aid users.

Expectations for Influenza, Zika, and MERS

H3N2 Influenza in this project is used as a proof of concept, as H3N2 HA influenza only reassorts and does not recombine. H3N2 Influenza is a seasonal, global disease where clades are defined by mutations from other strains, making it the most compatible with the Hamming Distance algorithm of the three diseases used to reduce the embeddings detailed in the methods. We use H3N2's HA sequences as they have a relatively high mutation rate compared to the other gene segments, it encodes a protein that is a target of human immunity, and has traditionally been used for analysis of influenza evolution. The genomes are 1701 bases long, with a mean bases missing of .04522 and median of 0. The population of H3N2 HA influenza used had a nucleotide diversity pi value of 0.0149, which is consistent with its short genome and low mutation rate. As these sequences are biologically relevant, short, and do not recombine, it can be assumed that H3N2 HA influenza is a good test case for Cartography. Zika in this project is used as

a test case. 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, while Influenza's clades were defined by mutations, zika's clades were defined by significant geographical introductions and outbreaks. Because of the difference in the definition of a clade, we used Zika to determine if the embeddings can not only recapitulate mutational but also geographical significance within its clustering. The genomes are 10769 bases long, with a mean bases missing of 913.613 and median of 154. The strains of MERS used had a nucleotide diversity pi value of 0.00535. With a longer genome and possible recombination, it can be reasonably assumed that Zika is a good test case for Cartography. MERS in this project is used as another test case. MERS is a recombinant virus, and because there are observed departures from strictly clonal evolution, it suggests that recombination is an issue for inferring MERS-CoV phylogenies. While its effect on human outbreaks is minimal, as humans are transient hosts with a smaller probability for co-infection, its effect is exacerbated in camel outbreaks. 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. The strains of MERS used had a nucleotide diversity pi value of 0.00235, which is more than double that of H3N2 HA Influenza. With a long genome, recombination, missing bases, and multiple hosts, it can be reasonably assumed that MERS is a good case to test and challenge Cartography.

Embedding clusters recapitulate phylogenetic clades for seasonal influenza A/H3N2

All four dimensionality reduction methods qualitatively recapitulated clade-level groupings observed in the phylogeny ((**fig:flu-embeddings?**)). 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 (light orange) and its subclade A2/re (dark orange) map to adjacent regions of all four embeddings. We observed the same pattern for A1 (blue-green) and its subclade A1a (light green-yellow) as well as for A1b (light green) and its subclades A1b/135K (light yellow) and A1b/135N (light orange). The clade 3c2.A (blue) and its subclade A3 (light red) 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.

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

To quantify the patterns we observed in (**fig:flu-embeddings?**), 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 ((**fig:flu-euclidean-vs-genetic-distance?**)). PCA and MDS provided the strongest linear mapping to genetic distance (Pearson's R2 = 0.693 and 0.684, respectively). This same mapping for the UMAP method was less of a linear function (Pearson's R2 = 0.378) 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 R2 = 0.280) 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 and UMAP provided the most accurate classifications (MCC = 0.75 and 0.67, respectively) and outperformed pairwise genetic distance (MCC = 0.60) and PCA (MCC = 0.67, (**fig:flu-within-and-between-group-distances?**)). 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.

Embedding clusters reveal outbreak and geographical patterns within Zika

All four dimensionality reduction methods recapitulated phylogenetic patterns observed in the phylogeny ((**fig:zika-embeddings?**)). PCA, after imputing missing data, had a similar structure to the findings in Metsky et.al., where the clades were clustered on a continuum of different clades instead of tightly clustered as seen in Influenza. Geographical introductions and outbreaks isolated from the others were placed at larger euclidean distances than related introductions.

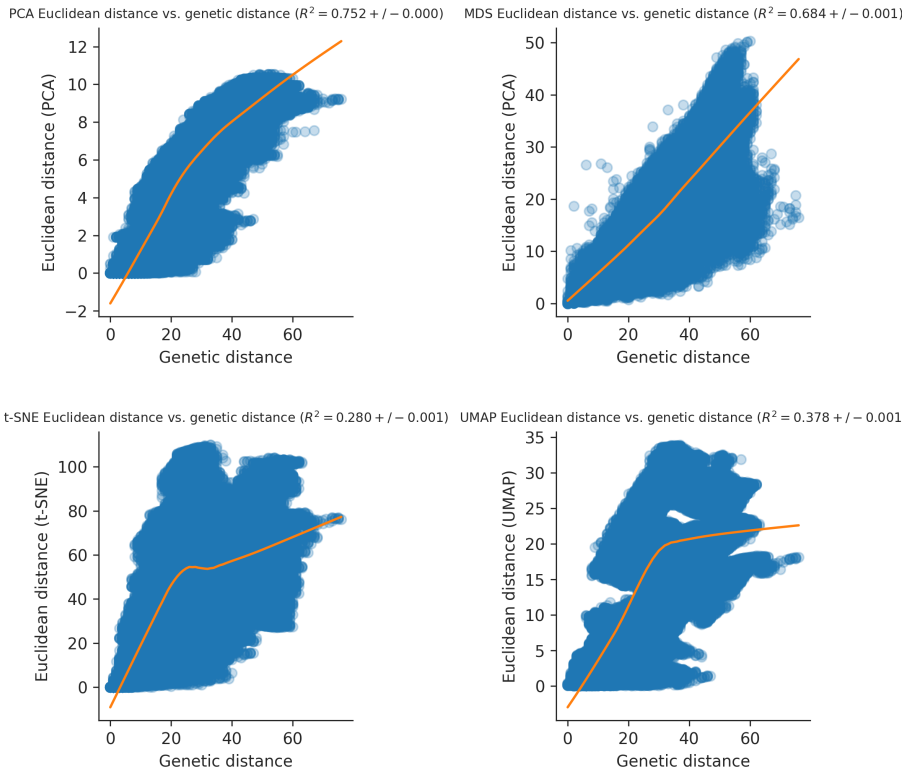


Figure 1: Mapping between Euclidean and genetic distances for all pairs of H3N2 strains by dimensionality reduction method.

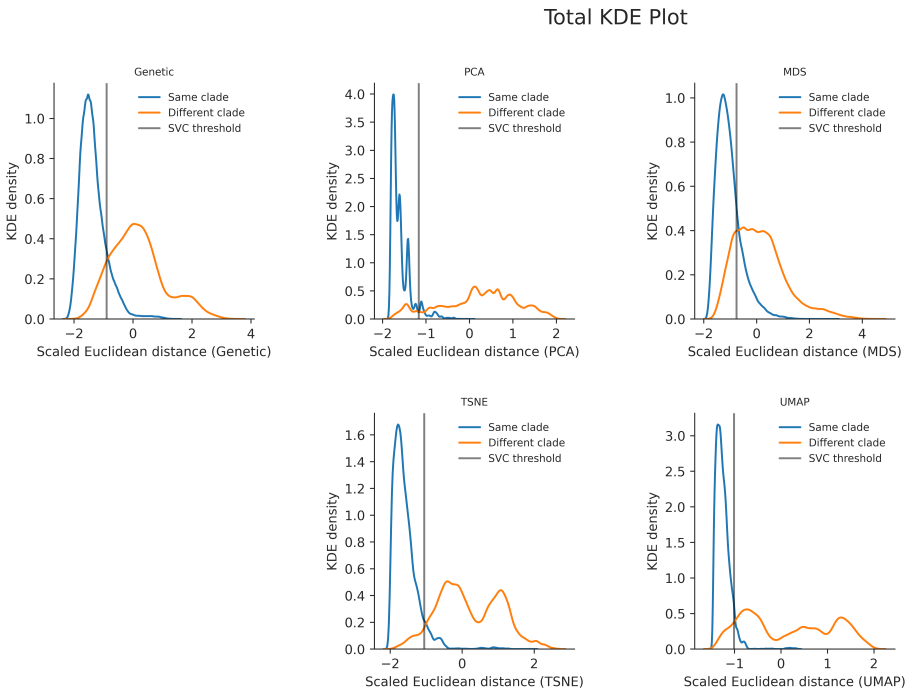


Figure 2: Distribution of scaled Euclidean distances between all pairs of H3N2 strains by clade status and dimensionality reduction method.

An example is clade c2 (light blue), an outbreak in Singapore and Thailand separated from the other geographical introductions in the Americas. Clade c10 (sky blue) 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 (Americas v Asia), and PC3 and PC4 are used to show the variance between clade c4 (light green) and c3 (blue-green) compared to clade c6 (orange) and c9 (red) (variance within the Americas). HDBSCAN performed on PC1 and PC2 was able to define 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 c3 was the parent of all Americas outbreaks, with clades c7, c5, c9, c6, and c4 being children of that outbreak. Therefore, clade c3 was split into multiple sections within each embedding, with these clusters correlating to different outbreaks within the parent clade c3 (the most notable in PC3 and PC4 and t-SNE). 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 the similarities 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 the embeddings are clustering by geographical introductions, not geography, which demonstrates the scope of this research in terms of variable clade definition. t-SNE and UMAP recognized a subset of a larger outbreak within clade c7 as by placing it farther away from the rest of the clade. These results suggest that genetic dissimilarity reduced via t-SNE and UMAP can be used to distinguish outbreaks from each other without imputation, an added benefit compared to PCA's sensitivity to missing data. Strain pairs were not as well separated in MDS space, but MDS did loosely cluster clades with genetically and evolutionarily distant clades farther away. This result suggests that MDS does better with un-imputed data than imputed, as the genetic distance normalization process is robust to gaps.

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

According to the Genetic vs Euclidean distance scatterplots, PCA, t-SNE, and MDS exhibited a piecewise linear relationship for pairs of strains that differed by no more than 50 nucleotides ((**fig:zika-euclidean-vs-genetic-distance?**)). For larger than a 50 nucleotide difference in genetic distance, PCA, t-SNE, and UMAP's LOESS line's slope is 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 R2 = .738 +/- .001). The UMAP mapping revealed two different clusters of points in the scatterplot, which is the stark Euclidean distance differences between clade c2 and the other strains due to its isolated sampling. This clustering is only seen in UMAP, revealing UMAP's sensitivity

to large amounts of outliers in the embeddings quality. t-SNE's mapping was fairly strong (Pearson's R2 = 0.522 +/- .002) and revealed that pairs of strains that differ by more than 50 nucleotides are unlikely to placed near each other in a t-SNE embedding.

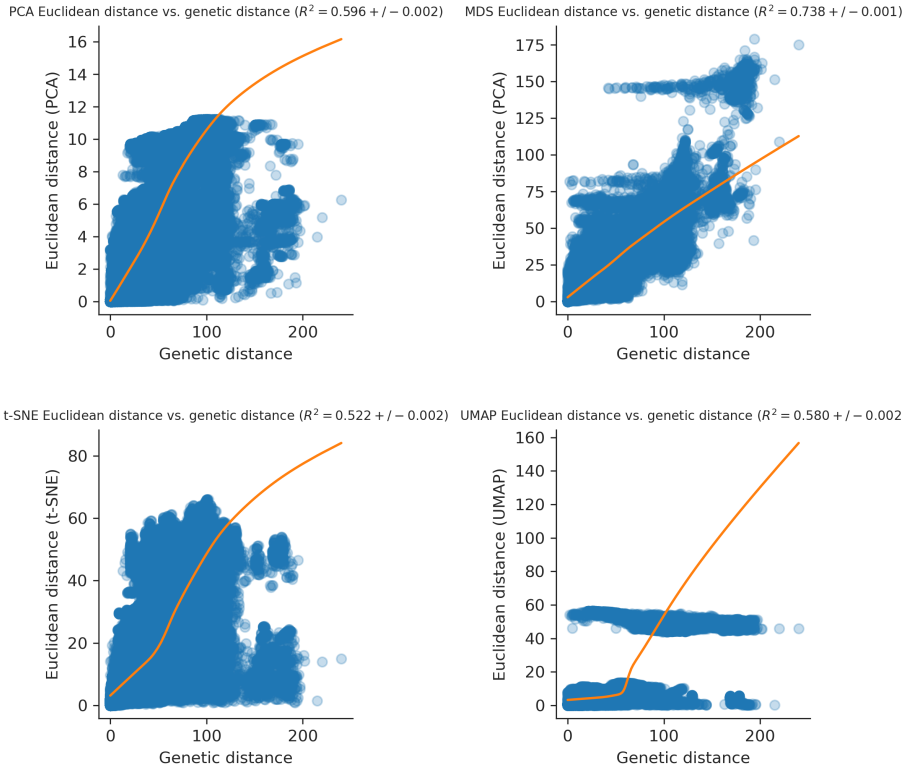


Figure 3: Mapping between Euclidean and genetic distances for all pairs of Zika strains by dimensionality reduction method.

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, (**fig:zika-within-and-between-group-distances?**)). 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 (FN: 7934 vs FP: 49397). MDS performed poorly (MCC = .41), confirming our expectation that MDS slightly underperforms genetic distance

based classification due to its emphasis on preserving global relationships. These results corroborate our previous conclusion about the potential benefits of using t-SNE embeddings for cluster analysis over genetic distance.

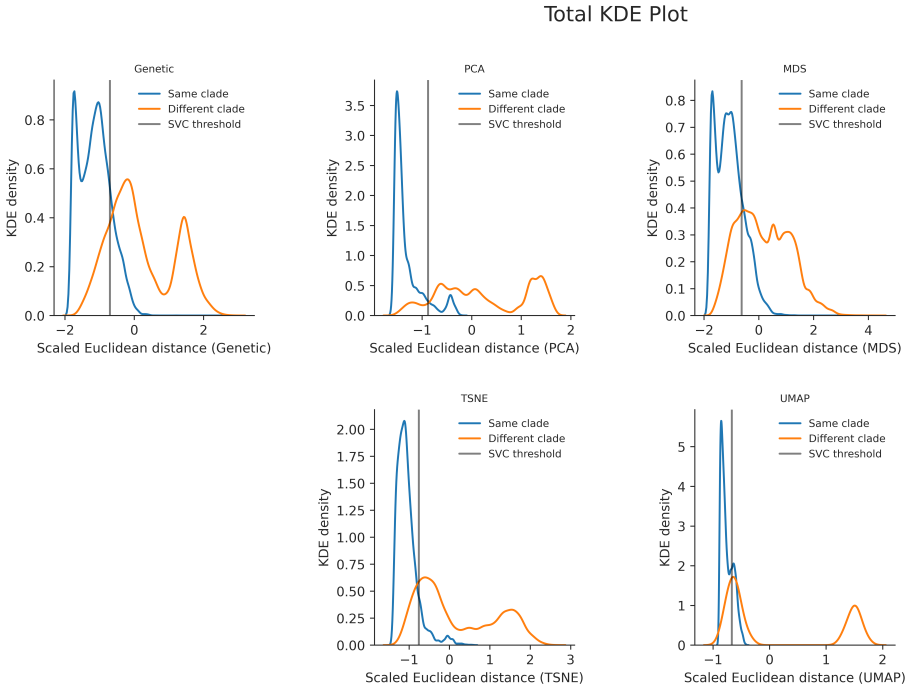


Figure 4: Distribution of scaled Euclidean distances between all pairs of Zika strains by clade status and dimensionality reduction method.

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 ((**fig:MERS-embeddings?**)). With MERS missing bases in multiple strains, imputation did not help add more depth to the alignment. To combat this issue, all strains missing more bases than 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 (clade_32), a human outbreak from Seoul and surrounding territories that is isolated from other clades, is a good example of an outlier with the most notable classifications in t-SNE and UMAP and to a lesser degree in MDS.

Because clades were defined as outbreaks sharing a common host, local clustering that revealed differences within a clade was expected; this divergence was seen in t-SNE and UMAP in Clade 13 (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, putting human host disease strains without a clade membership in the nearest related clade. The "other" clade membership strains are inferred to be direct inter-host infections, so between host clustering recapitulates the genetic similarity between these inter-host strains. t-SNE performed exceedingly better than the other embeddings at differentiating between intra-host clades highly related in the embedding, outperforming UMAP and MDS. 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), which were not separated in UMAP. This suggests that t-SNE is a stronger tool for viewing genetically homogeneous populations.

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

According to the Genetic vs Euclidean distance scatterplots, 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 ((**fig:MERS-euclidean-vs-genetic-distance?**)). 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. While these embeddings are still using local patterns to map genetically distant strain combinations at higher and lower euclidean distances, t-SNE and UMAP select a low or high euclidean distance depending on the strain relationship. This can be attributed to the two larger clusters in both the UMAP and t-SNE embeddings, with Clades 27 through 32 in one group, and the rest in the other. MDS provided the strongest linear mapping to genetic distance (Pearson's R2 = 0.759 +/- .003). UMAP and t-SNE's scatterplots shared many similar characteristics in terms of shape, spread, and clustering (Pearson's R2 = 0.203 +/- .005 and 0.253 +/- .005 respectively). The same mapping for PCA was incredibly weak and nonlinear (Pearson's R2 = 0.023 +/- .001).

Just as in Influenza and Zika, t-SNE provided the most accurate classification of the embeddings (MCC = .625) but did not outperform pairwise genetic distance (MCC = 0.71, (**fig:MERS-within-and-between-group-distances?**)). 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 = .479), 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 (FN: 309 vs FP: 2164). MDS performed much better than in Influenza

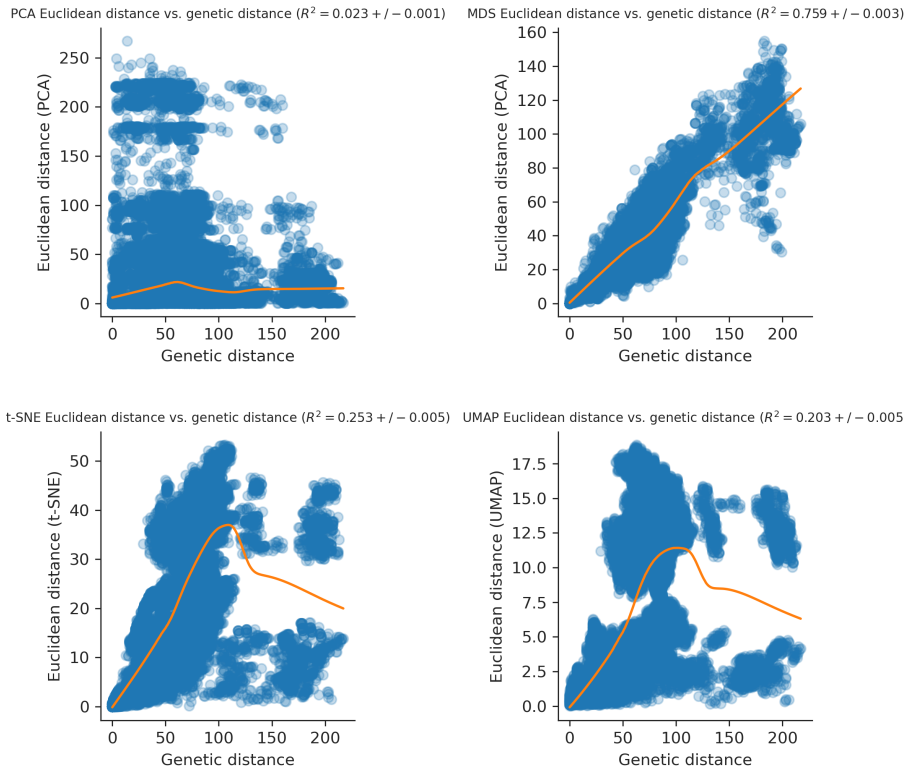


Figure 5: Mapping between Euclidean and genetic distances for all pairs of MERS strains by dimensionality reduction method.

and Zika (MCC = .624), and with MDS being the most global reduction of the data and genetic distance outperforming all the embeddings, shows MERS to be a disease best classified without local patterns. These results corroborate our conclusion about using t-SNE embeddings for cluster analysis, but suggest viewing and quantifying the data through multiple reductions in order to create the best view of the data.

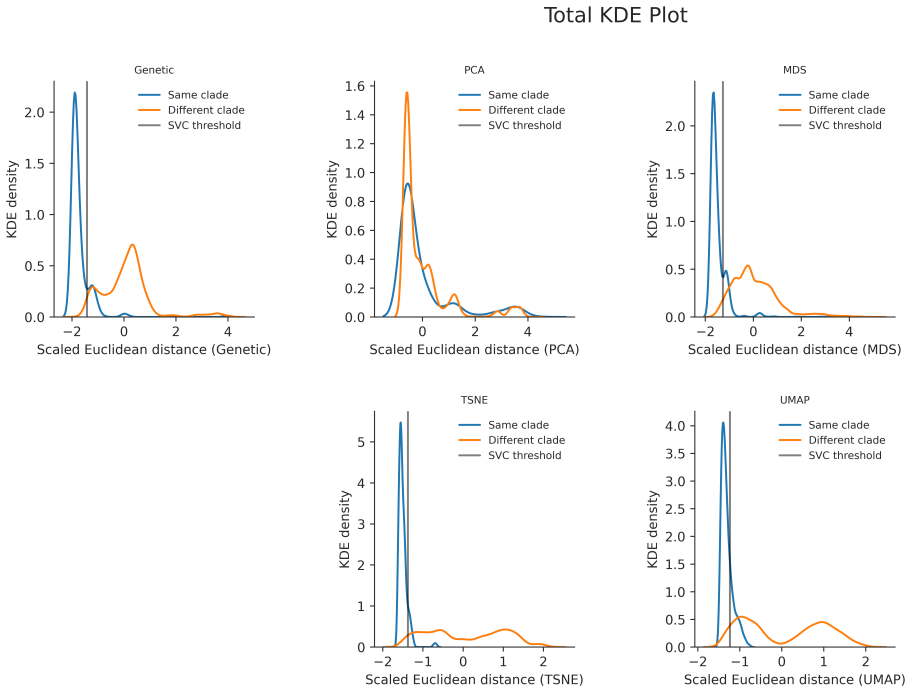


Figure 6: Distribution of scaled Euclidean distances between all pairs of MERS strains by clade status and dimensionality reduction method.

Discussion

In this paper, we analyzed the usage of PCA, MDS, t-SNE, and UMAP to better understand population structure in varying types of disease with differing clade assignments. We accomplished this both qualitatively 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 Cartography's public automated code pipelines and scripts, more accessible to scientists and the public.

The diseases we used to analyze varied in genome length, genetic variance,

and phylogenetic clade assignment due to different modes of transmission and mutability. Across all the diseases, we found PCA to be incredibly sensitive to missing data. While this is not a major issue in smaller genomes with less variance within the population, imputing missing data and dropping strains to create a useful PCA embedding introduced noise into the embeddings. We attributed this drawback to the larger flaw of using sites on a genome as features to find useful patterns. This, however, does make PCA useful in pulling out low quality and misnamed strains in a population. A future direction would be defining a euclidean threshold between low quality and normal strains to find outliers within a population. As it is common practice to run flu builds all the way through to auspice JSONs, view them, and find a handful of outliers that should have been 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.

In terms of local patterns and clade detection, t-SNE consistently did better than the other embeddings. An advantage of t-SNE over PCA is its robustness to missing data while preserving similar, if not better, quality results. t-SNE performed the best at finding outbreaks, newly created clades, and local patterns within larger clades. A question we aimed to answer with this research was the level of their phylogeny that clusters within each embedding were being developed. Using the community builds of Cartography's phylogenies colored by t-sne x (component 1), we analyzed Flu, Zika, and MERS clade structure and coloring. We found that within Flu, t-SNE clusters exactly as defined by the clades, which were defined by genomic variance, and we attributed this success to the influenza population being constituted of small genomes with well defined clades. Within Zika, 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. Within MERS, the clades were defined by outbreaks per host, and t-SNE did the best at pulling out these per host outbreaks of the four embeddings, correctly identifying outbreaks with less than 6 samples and 3 nodes distance in the phylogenetic tree. Beyond qualitative patterns, t-SNE consistently outperformed most of the embeddings and genetic distance for classification of between and within clade relationships, while genetic did better than t-SNE for MERS. As genetic variance and genome length continues to increase, t-SNE continues to work well, but the added noise does create less segmented embeddings. Of the four embeddings and genetic distance, the best indicator for the mathematical determination of clades (how many relationships are preserved) is t-SNE's euclidean distance and genetic distance. A future direction for testing the applicability of t-SNE would be a cross-validation test, where a t-SNE threshold of euclidean distance between relationships is used on a population of the same disease from some years later. This could potentially help determine the usefulness of a t-SNE's euclidean distance based threshold in understanding future trends in present populations. 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. UMAP consistently performed worse with classification and finding local patterns, and we conclude that this is due to UMAP's inability to cluster finer patterns in genetically variable populations, and instead creates larger groups of often multiple clades and outbreaks.

Our recommendations for algorithm choices have to do with the many factors that create a population dataset. For a population with smaller genomes and very few missing bases, PCA and t-SNE work best at both defining clades, and embeddings that convey useful information. For genomes with around 10K bases and more missing data, PCA will still work with basic imputation (the caveat, however, is the added noise). To get around this issue, t-SNE creates a similar, if not 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 of genetic distance. For large genomes (30K bases+), lots of missing data with fewer samples, and diversity, t-SNE is the most useful of the embeddings, while it is not as strong as it is in smaller, less variable genomes. Genetic distance works best in these populations for classifying relationships as within or between clade. A visualization of the raw pairwise data would be achieved through MDS, as MDS consistently created the strongest correlated linear relationship between pairwise and euclidean distance.

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 opened a gateway for its usage in practical applications in the future. This paper has delved deeper into the scope of these embeddings, and created tools as used for this paper that are public, easy to use on other datasets, and will be on boarded into nextstrain-augur for ease of use. The hope is that scientists will now be able to use these embeddings to further understand their dataset and diseases, and use genetic cartography with rapidly emerging diseases such as SARS-CoV-2, where a lack of data and samples can make traditional epidemiological tools less accurate.

Materials and Methods

The analysis environment can be recreated using conda and all installation instructions are available on Cartography'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 as well as by reasonable phylogenetic signal in terms of mutations on branches. 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 created, 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 in order to get a better embedding result. This was only applied to PCA, as the hamming distance algorithm disregards missing bases. Imputation was tested for MERS, but due to entire columns of missing data for MERS, we had to drop 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 that have 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 mean complete genetic dissimilarity, which is what counting gaps implied.

We reduced the similarity distance matrix through MDS, t-SNE, and UMAP, plotted using Altair ,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. 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 through an exhaustive grid search, which picked the best values by maximizing Matthews Correlation Coefficient on the confusion matrix created from a Supported Vector Machine's classification. UMAP's minimum distance and nearest neighbors were tuned, and t-SNEs perplexity and learning rate were tuned as well. 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. The proportion value was used to determine the nearest neighbors value for the UMAP plots per disease. 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 will 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 rather than look, and a splayed structure points to local structure preservation over global. 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.

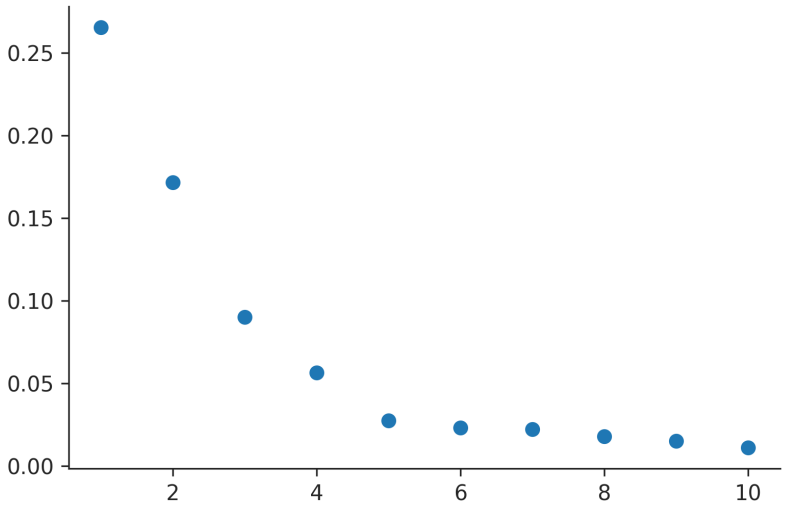
Between vs Within clade KDE Density Plots:

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

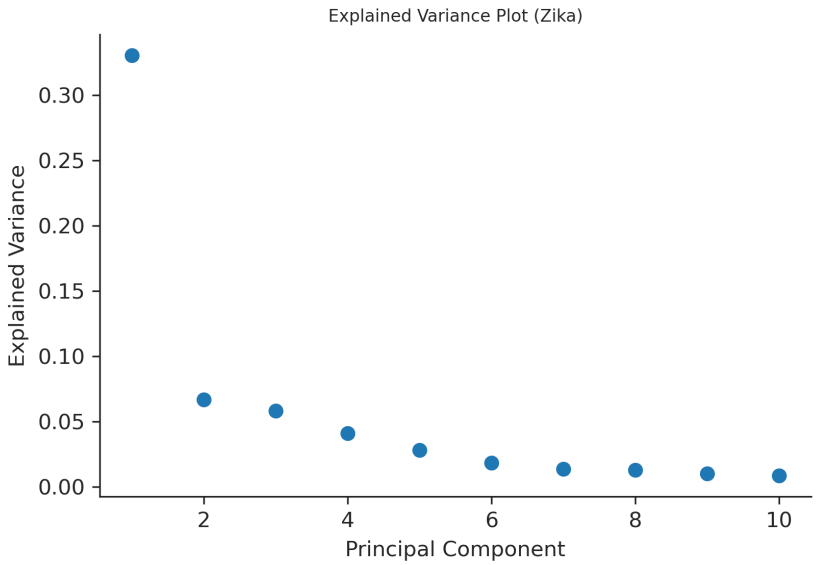
to classify a given pair of genomes as within vs between clades. The larger the median ratio between the two curves presented per clade relationship, the higher the relative probability that the embedding will accurately predict if two strains with any specific distance is a between or within clade relationship. 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.

Supplementary Figures and Analysis

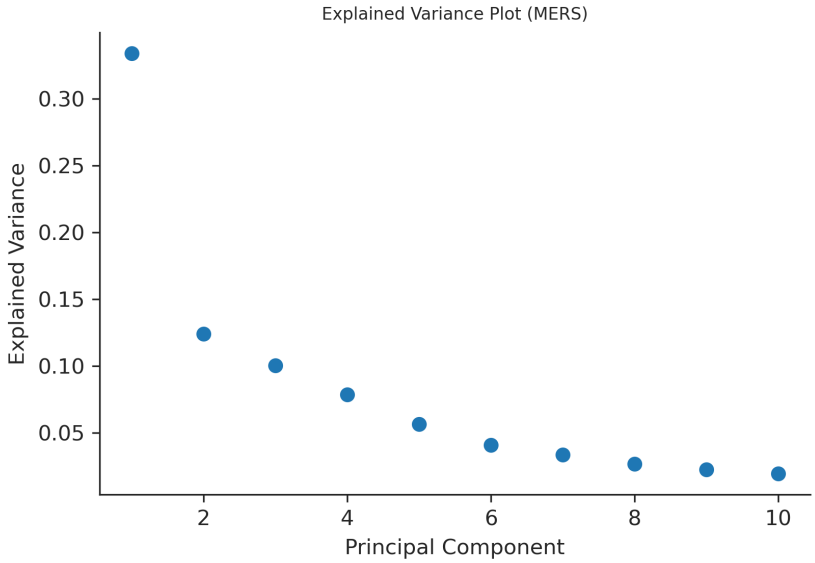
Explained Variance Plots for PCA



Flu



Zika



MERS

PCA Full Plots

Flu

Zika

MERS

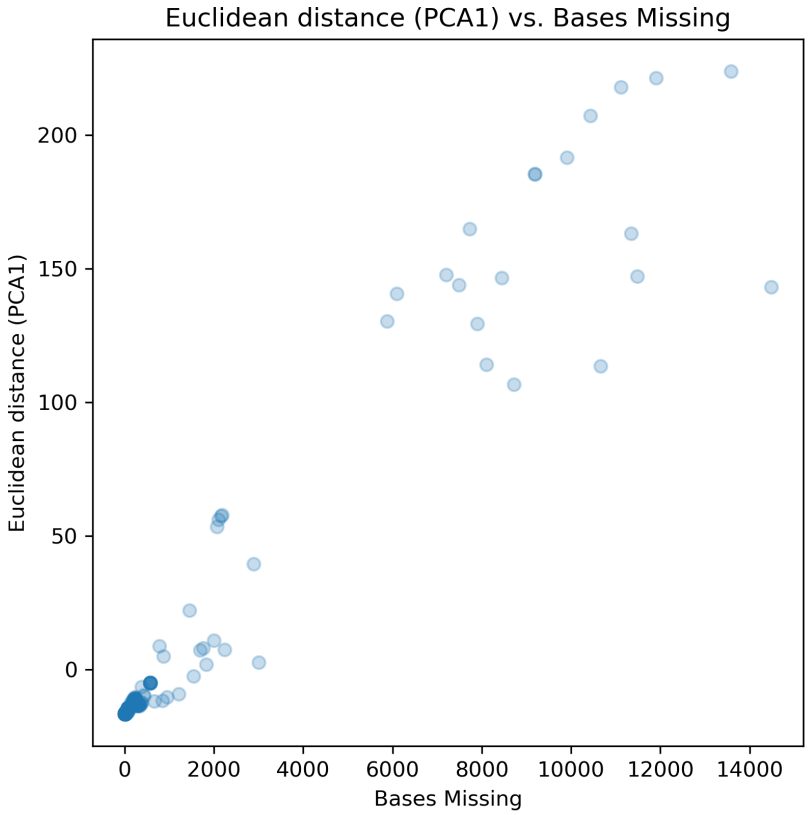
MDS Full Plot:

Flu

Zika

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.