Cartography

Sravani Nanduri

DESCRIPTION OF THE PROBLEM:

Phylogenetic inference is a fundamental tool for understanding genealogical relationships among human pathogenic viruses. However, recombination and reassortment of viral genomes 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 in the data (Posada and Crandall 2001) (Martin et al. 2017). An alternate strategy is to compare viral genomes with alternative methods that do not make the same strong assumptions as phylogenetic inference (e.g., PCA, MDS, etc.). PCA has been used to estimate and model ancestry in unrelated individuals and plot peoples genomes to reveal patterns in national origin (Alexander, Novembre, and Lange 2009) (Novembre et al. 2008), and was also used to genotype major classes of structural variants - structural differences such as deletion, duplication, and novel sequence insertion - in diverse populations to map their population stratification (Peter H. Sudmant 2015). PCA was also used to reveal Zika's genetic diversity and spread in the Americas by assessing 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, for the difference that Rambaut et .al. 2008 looks at between-gene diversity rather than within-gene (Rambaut et al. 2008). The MDS analyses showed tight clustering between genes, which suggested 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 better capture both discrete and continuous patterns of variation in human

genomes across a genetic continuum, and the embeddings were able to show relationships between genotype, phenotype, and geography (Diaz-Papkovich et al. 2019). While Diaz-Papkovich et.al. and Metsky et.al. 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 investigate the degree to which this method can recapitulate known phylogenetic relationships for viruses whose genomes are phylogenetically tractable (we used influenza h3n2 HA and Zika, with of evolutionary rates of around 4.04 × 10ˆ-4 (95% HPD: 1.32 × 10ˆ-4, -7.41 * 10ˆ-4) and 9.57 * 10ˆ-4 (95% Highest Posterior Density: 8.28 - 10.9 * 10ˆ-4) subs/site/year respectively). We apply this method to viruses whose genomes are known to undergo substantial recombination, such as MERS (evolutionary rate of around 5.15×10ˆ-3 (HPD 4.62 * 10ˆ-3, -5.70 * 10ˆ-3) substitutions/site/year).) and SARS-CoV-2 to assess how well each method is able to reconstruct previously identified biologically-meaningful clusters.

Recombination: occurs when at least two viral genomes co-infect the same host cell and exchange genetic segments. Shuffling/reassortment, a particular type of recombination, occurs in viruses with segmented genomes, which by interchanging complete genome segments, gives rise to new segment combinations (Pérez-Losada et al. 2015).

METHODS:

Materials:

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

Methods:

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, which was split into a Newick tree and Aligned FASTA file. (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 taken from the newick tree created in Gytis' and

Bedford's paper (Dudas et al. 2018). 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.

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) 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 for the confusion matrix created from a Supported Vector Machine splitting the between vs within clade KDE density plots. 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. For example, the more nearest neighbors, the higher the computational load, and while smaller minimum distances can break

connectivity between clusters, they will not change the groupings of individuals.

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's 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. 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 local patterns rather than global, and a splayed structure points to global structure preservation over local. Therefore, the closer the Pearson Coefficient is to 1, the better the embedding is at preserving genetic dissimilarity 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 for more information.

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.

Cross Validation:

To quantify the patterns seen in the plots and further understand the benefits and drawbacks to each embedding method, we ran a cross validation analysis using the h3n2 data from 2016-2020. This test would answer if viral genomes from the same clades have smaller Euclidean distances in a given embedding than viral genomes from different clades. It also answers if any of these embeddings classify pairs of viral genomes better than genetic Hamming distance, which would help decide if creating the computationally intensive embeddings is worthwhile.

We constructed the time-series cross validation test by creating a K-fold cross-validation generator, with 5 folds containing both training and validation strains for data from 2016 to 2018. A KDE Density plot was created per training fold. A Support Vector Machine optimized a distance threshold that most accurately predicted clade relationships for that given fold, and a confusion matrix was created per fold on the validation data. The distance threshold values given by the SVM were pooled via an average, and one final confusion matrix was created on the 2018-2020 h3n2 data to not counfound the results.

Classification Accuracy Test

For MERS, a classification accuracy test was run to determine the accuracy of using just the embeddings to determine the clade the strain is from in the Newick tree. This would allow us to further understand how the embeddings could be used in recombinant diseases, where the host would not matter as much as the genetic simmilarity and different between strains. Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBSCAN) was used to find clusters in the embeddings without any set number of clusters. The Newick Tree predefined clades were used as the "Truth", and the labels assigned by HDBSCAN were compared against these values. The accuracy value was determined by the proportion of correct relationships retained to total relationships.

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, does not affect the scale of the data, and does not normalize the data as part of the algorithm. PCA preserves long range distances but hides finer-scale details. Because PCA is almost entirely focused on retaining the global structure and variance of the data, and one of its limitations is revealing patterns locally. PCA is not an algorithm to be used on a similarity matrix, and is instead intended for transformed and normalized multidimensional data. In the context of this paper, PCA will be used on transformed and normalized genetic data and not on the similarity matrix described above.

Multidimensional Scaling (MDS) refers to statistical techniques that reduce the complexity of a data set by quantifying similarity judgments, which increases 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 preserves global patterns over local, but the algorithm's importance on translating dissimilarity to distance does preserve some larger patterns in local structure as well. In the context of this paper, MDS will cluster data into sparse

"sections" of the map while not creating actual clusters.

t-distributed Stochastic Neighbor Embedding (t-SNE) visualizes high-dimensional data by giving each datapoint a location in a 2 to 3 dimensional map. t-SNE is focused largely on local structure 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 heavily on projecting similarity rather than dissimilarity (Maaten and Hinton 2008). Because t-SNE reduces data's dimensionality based on local properties of data, data with intrinsically high dimensional structure will not be projected accurately. In the context of this paper, t-SNE will create tight clusters that clearly indicate genetic similarity, but will not create an accurate global picture of the data.

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 a patchwork of low-dimensional representations of neighbourhoods that groups genetically similar strains together on a local scale while better preserving long-range topological connections to more distantly related strains. Some limitations include its lack of maturity - this novel technique does not have firmly established or robust practices and libraries to use UMAP best. In the context of this paper, UMAP will reveal a tightly clustered set of data that retains both the global structure of the data and the clusters and similarities present at the local level.

Influenza:

h3n2 Influenza in this project is used as a proof of concept as h3n2 HA influenza only reassorts and does not recombine. The genomes are 1701 bases long, with a mean bases missing of .045217 and median of 0. The evolutionary rate for H3N2 HA influenza from Baysian structured coalescent estimate is 5.15×10^{-3} (HPD 4.62 * 10ˆ-3, -5.70 * 10ˆ-3) substitutions per site per year. 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. As these sequences are biologically relevant, short, and do not recombine, the genomes can be reasonably assigned to phylogenetic clades. Therefore, it can be assumed that h3n2 HA influenza is a good test case for 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 (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 (orange) and its subclade A2/re (red) map to adjacent regions of all

four embeddings. We observed the same pattern for A1 (purple) and its subclade A1a (pink) as well as for A1b (brown) and its subclades A1b/135K (gray) and A1b/135N (blue). The clade 3c2.A (red) and its subclade A3 (light blue) 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 UMAP provided the strongest linear mapping to genetic distance (Pearson's R2 = 0.693 and 0.615, respectively). This same mapping for the MDS method was less of a linear function (Pearson's R2 = 0.468) than a piecewise function of two parts. Strain pairs with fewer than 30 nucleotide differences were not as well separated in MDS space as strains with greater genetic distances. This result suggests that MDS might be most effective for distinguishing between more distantly related strain pairs. t-SNE's mapping was the weakest (Pearson's R2 = 0.269) 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 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 classifiction 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 PCA provided the most accurate classifications (MCC = 0.73 and 0.68, respectively) and outperformed pairwise genetic distance (MCC = 0.65) and UMAP (MCC = 0.63, Figure 3). MDS performed poorly (MCC = 0.41), confirming our expectations based on MDS's piecewise linear relationship with genetic distances. 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.

SUMMARY OF RESULTS FOR INFLUENZA

PCA (within clade -1.539, between clade -.058, MCC: DNE accuracy:.888)

MDS (within clade -1.361, between clade median 0.108, MCC: 0.560 accuracy:0.922)

TSNE(within clade median -1.694, between clade median 0.186, MCC: 0.784 accuracy: 0.958)

UMAP(within clade median -1.337, between clade median 0.034, MCC: 0.631, accuracy: 0.926)

PCA: We observed visually identifiable clusters within the data (Figure 1B), and found less distance between genetically and geographically similar clades (3c2.A, A1, A1a, and A1b), and divergence between other clades (3c3.A). PCA is consistent with phylogenetic observations; this is reflected in the points' constant distances from the LOESS line in the Euclidean and Hamming distance scatterplot for PCA (Figure 2A). The LOESS line with a Pearson Coefficient of .693 was fairly linear, which upholds preexisting beliefs about the algorithm. Euclidean distance can be used with some confidence to distinguish strains by clade status, as the density of within clade relationships is concentrated at lower distances (within clade median of -1.539) than the density of the between clade relationships(between clade median of -0.058) (Figure 3A) The threshold distance value calculated using the SVM run on PCA gave a confusion matrix with an accuracy of 0.888, which corroborates that Euclidean distance is a fairly strong indicator of clade status.

MDS: While we did observe visually identifiable clusters, we found many clades overlapped each other, particularly clades A2/re, A1, A2, A1b/135N, and A1b/135K (Figure 1C). MDS is fairly consistent with phylogenetic obervations; this is seen in the points's constant and short distance from the LOESS line in the Euclidean and Hamming distance scatterplot (Figure 2B). The LOESS line with Pearson Coefficient of .468 is fairly linear, with a steeper slope starting from a genetic distance of 30; this upholds preexisting assumptions about this method. Euclidean distance can be used with some confidence tp distinguish strains by clade status, as the bulk of within clade relationships are defined from -2 to 0 (within clade median of -1.361) while the between clade was at 0 and above (between clade median of 0.108) (Figure 3B). The threshold distance value calculated using the SVM run on MDS gave a confusion matrix with with a Matthews Correlation Coefficient of 0.560 and an accuracy of 0.922, which corroborates that Euclidean distance is fairly useful as an indicator of clade status.

t-SNE: We observed t-SNE going beyond compartementalizing by clade and actually revealing a stronger understanding of hierarchical structure and resemblance to the tree produced for the data. In particular, t-SNE split A1b into three clusters, which were lineages directly or incredibly related to each other, revealing a tuning to local patterns not seen in PCA and MDS; this is corroborated by the points' splaying out from the LOESS line as genetic distance increases in the Euclidean and Hamming distance scatterplot for t-SNE (Figure 2C). The LOESS line had a Pearson Coefficient of .269 due to this splay and

overall tuning to local structure. Euclidean distance can be used with confidence to distinguish strains by clade status, with a within clade median of -1.694 and between clade median of 0.186 in the KDE density plot for t-SNE's clade statuses(Figure 3C). While the difference is smaller than PCA and MDS, the KDE plots for those embeddings were skewed by the "outlier" clades, which PCA placed very far away from the other clades. The threshold distance value calculated using the SVM run on t-SNE gave a confusion matrix with with a Matthews Correlation Coefficient of 0.784 and an accuracy of 0.958, corroborates that Euclidean distance is useful as an indicatior of clade status in spite of of t-SNE splitting clades due to local diversity.

UMAP: We observed UMAP placing tightly clustering similar datapoints, which we found can make it difficult to view hierarchical structure as seen in t-SNE. In the Euclidean vs Hamming distance scatterplot, we observed the data points splaying out from the LOESS line in lower genetic distances, but the more distinguishing characteristic was the clustering at lower versus higher euclidean distances(Figure 2D), which upholds preexisting beliefs about the locality of manifold reduction techniques such as UMAP. The correlation of the data points to the LOESS line is described by UMAP's Pearson Coefficient of .615. Euclidean distance can be used with confidence to distinguish strains by clade status, with the bulk of within clade relationships are defined from -2 to -1 (within clade median of -1.337) and between clade at 0 and above (between clade median of 0.034). The threshold distance value calculated using the SVM run on MDS gave a confusion matrix with with a Matthews Correlation Coefficient of 0.631 and an accuracy of 0.926, which reveals that Euclidean distance is useful as an indicator of clade status.

Figure One

Figure Two

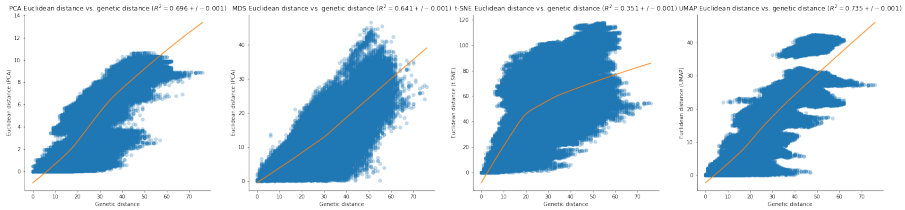
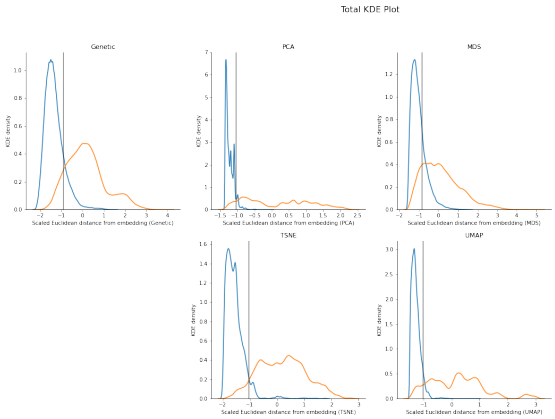


Figure Three



Zika:

Zika: 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, Zika is better compartamentalized by region than by clade - this is why the clades determined in the analysis for Zika were lagely based on important geographical introductions. The evolutionary rate from Baysian structured coalescent estimate is 4.04 $\times$ 10ˆ-4 substitution/site/year (95% HPD: 1.32 $\times$ 10ˆ-4, -7.41 * 10ˆ-4) substitutions per site per year. The genomes are 10769 bases long, with a mean bases missing of 913.613 and median of 154. With a longer genome and reassortment, it can be reasonably assumed that zika is a good test case for Cartography.

All four dimensionality reduction methods qualitatively recapitulated clade-level groupings observed in the phylogeny (Figure 1). PCA, after imputing missing data, had a similar global structure to the findings in Metsky et.al., where the clades were featured on a continuum of shifting between clades instead of tightly clustered as seen in Influenza. Strains from the same clade were tightly grouped in t-SNE and UMAP, and more loosely clustered in the MDS embedding. Closely related clades still tended to tightly cluster in MDS, UMAP, and, to a lesser extent, t-SNE. In all four embeddings, clades that were genetically and evolutionarily divergent from the other clades were incredibly distant in embedding space. For example, the clade c2 (red) and the other clades are incredibly distant in UMAP and t-SNE, and all the other clades cluster much closer together. in MDS, however, clade c2 (red) is not placed at a large distance from the other clades, which is surprising considering the MDS embedding for Influenza. The clade c5 (yellow) and its most related clade c7 (pink) clustered tightly in all embeddings. Clade c4 (green) was somewhat arbitrarily split into

two clusters in the UMAP embedding. According to the phylogenetic tree, there is no internal node split between the two clusters in clade c4, which is not the expectation for UMAP.

According to the Genetic vs Euclidean distance scatterplots, all four methods exhibited a consistent linear relationship for pairs of strains that differed by no more than 50 nucleotides (Figure 2). After 50 nucleotide difference in genetic distance, PCA, t-SNE, and UMAP increase much faster in a piecewise fashion, revealing that these embeddings are using local patterns to map genetically distance strain combinations farther away for better visualization. This is the expectation for t-SNE and UMAP, but is surprising to see in PCA. PCA and UMAP provided the strongest linear mapping to genetic distance (Pearson's R2 = 0.573 +/- .002 and 0.580 +/- .002, respectively). The UMAP mapping revealed two different clusters of points in the scatterplot, which is the stark Euclidean distance differences between clade c2 (red) and the other strains. This same mapping for MDS was less linear and the weakest (Pearson's R2 = 0.253 +/- .002). Strain pairs were not as well separated in MDS space, but MDS did loosely cluster clades with genetically and evolutionarily distant clades farther away (Clade c1, the strains dated in 2013-2014). This result suggests that MDS does better with unimputed data than imputed, as the genetic distance normalization process is robust to gaps. t-SNE's mapping was fairly strong (Pearson's R2 = 0.522 +/- .002) and revealed that only closely related strains map near each other in t-SNE space. Pairs of strains that differ by more than 50 nucleotides are unlikely to placed near each other in a t-SNE embedding.

Once again, t-SNE and PCA provided the most accurate classifications (MCC = 0.61 and 0.56, respectively) and outperformed pairwise genetic distance (MCC = 0.49) and UMAP (MCC = 0.27, Figure 3). 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: 28369 vs FP: 970). MDS performed poorly (MCC = ???), confirming our expectations based on MDS's piecewise linear relationship with genetic distances. These results corroborate our previous conclusion about the potential benefits of using t-SNE embeddings for cluster analysis over the computationally simpler genetic distance, but it also affirmed that t-SNE does reveal global relationships when the genetic distance is incredibly divergent.

SUMMARY OF RESULTS FOR ZIKA:

PCA (within clade -.421, between clade -.214, MCC: DNE accuracy: .855)

MDS (within clade -1.179, between clade median -.083, MCC: .541 accuracy:0.898)

TSNE(within clade median -1.142, between clade median -.312, MCC: 0.616 accuracy: 0.913)

UMAP(within clade median -.783, between clade median -.544, MCC: 0.368, accuracy: 0.877)

PCA: The PCA plot for Zika colored by clade (Figure 4B) did not reveal any visually identifiable clusters in the data, and the data's overall structure also did not reveal anything interesting. Visually, there was no relationship between Hamming Distance and Euclidean Distance within the PCA embedding, shown by the completely random placement of points on the scatterplot (Figure 5A). To quantify this visual observation, the Pearson Coefficient was .089, which reveals close to no correlation. The lack of visible clustering in this embedding reveals that scaling and centering nucleotide data does not capture the genetic diversity present between each genome.
Euclidean distance does not help distinguish viral genomes by genetic diversity (Figure 6A). In the KDE density plot for PCA, the within and between KDE density curves overlap completely, and there is no place the SVM can correctly optimize for clade status, which is revealed through the within clade median of -0.421 and the between clade median of -0.214. The threshold distance value calculated using the SVM run on MDS gave a confusion matrix with an accuracy of .855, which is most likely skewed due to the large dataset and size based accuracy value.

MDS: The embedding seemed to differentiate between the clusters in the data on a very global scale (Figure 4C). There were a few visually identifiable clusters in the data, such as clusters containing clades 1, 5, and 4, but clades such as 7 and 9 were almost completely overlapped by other clades. Because MDS tends to reveal global structure over local, MDS tries to as accurately as possible preserve the data's overall structure, which means it doesn't reveal many local patterns. In the Euclidean and Hamming distance scatterplot, the points begin relatively close to the LOESS curve, but begin to move farther and farther away as the genetic distance increases, and this divergence is corroborated by the Pearson Coefficient of .199 for the plot. The LOESS line plateaus at a genetic distance of 50, revealing that as points become more and more divergent, MDS places them at similar euclidean distances, which is a hallmark of algorithms that preserve some local structure. Euclidean distance does a fairly accurate job distinguishing viruses from similar and different clades (Figure 6B). In Figure 6B, the within clade relationship curve on the KDE density plot is at lower euclidean distances than the between clade relationships, which there is a lot of overlap. The embedding had the within clade median of -1.179 and the between clade median of -0.083, which reveals that it's fairly good at compartementalizing the data by euclidean distance. The threshold distance value calculated using the SVM run on MDS gave a confusion matrix with with a Matthews Correlation Coefficient of 0.541 and an accuracy of 0.898, which reveals an embedding where the inferred clade statuses are fairly accurate given no other information.

t-SNE: This embedding does a very good job of clustering the data; every clade is a visually identifiable cluster, and clades more similarly related genetically are closer together in the plot (Figure 4D). The data points from different clades

did not overlap each other much, and the embedding did revealed different relationships than the rendering of the tree, giving less euclidean space between points that were more genetically diverse and creating a "threshold" for how much genetic diversity would impact euclidean distance. In the euclidean and Hamming distance scatterplot for t-SNE (Figure 5C), the points splay out from the LOESS line as genetic distance increases, which is expected for this algorithm. This correlation can be expressed through the scatterplot's Pearson Coefficient of .499. The decrease in slope of the LOESS line after a genetic distance of 100 usually points to an embedding that gives little importance to the exact distance between clusters and instead focuses on cluster shape and spread. These statistics reveal that t-SNE gives a large focus to finding new patterns within the data while still revealing similar patterns to the tree. It is quite easy to distinguish viruses of different and same clades given a euclidean distance. In the KDE Density plot for t-SNE, the bulk of the same clade relationships fall to the left of the between clade relationships, which reveals a well clustered and compartmentalized embedding of these data points (Figure 6C). To quantitatively reveal this pattern, the embedding had a within clade median of -1.142 and the between clade median of -0.132, which is lower because of the "splay" and local patterns t-SNE tends to optimize for. The threshold distance value calculated using the SVM run on t-SNE gave a confusion matrix with with a Matthews Correlation Coefficient of 0.616 and an accuracy of 0.913, which reveals an embedding where the inferred clade statuses are fairly accurate given no other information.

UMAP: The embedding did translate the data differently than the rendering of the tree, where genetically similar strains were very densely packed and genetically different clades were incredibly far apart in UMAP euclidean space (Figure 4E). There were visually identifiable clusters, but the distance disparities made it hard to separate clades 9, 7, 6, and 3 from each other due to how close the clusters were. The pattern can be seen and explained in the Hamming and genetic distance scatterplot for UMAP (Figure 5D). In UMAP's scatterplot, there are two clusters of data points on the LOESS line, instead of the equal spread of points over euclidean and genetic space seen in the other 3 embeddings. These clusters point to a local pattern preserving algorithm that places genetically similar strains incredibly close together - a Euclidean distance around 0 to 20 - but places genetically different strains almost 50 to 80 apart. The LOESS line reflects this increase through its exponential-like growth, and the density of these two clusters can be quantitatively shown through its Pearson Coefficient of .595. Euclidean distance is not a very strong measure to distinguish viruses from the same or different clades. In the UMAP KDE Density plot (Figure 6D), the clades are all places close together save for clade 2, which is incredibly divergent compared to the others. Because all other between clade distances are the same as within clade distances, UMAP had a within clade median of -0.783 and the between clade median of -0.544, meaning it is not easy to optimize an SVM with euclidean distance for clade relationships. The SVM statistics corroborate this, with the threshold distance value calculated using the SVM run on UMAP

giving a confusion matrix with with a Matthews Correlation Coefficient of 0.368 and an accuracy of 0.877, which reveals an embedding where the inferred clade statuses are fairly accurate given no other information.

Figure Four

Figure Five

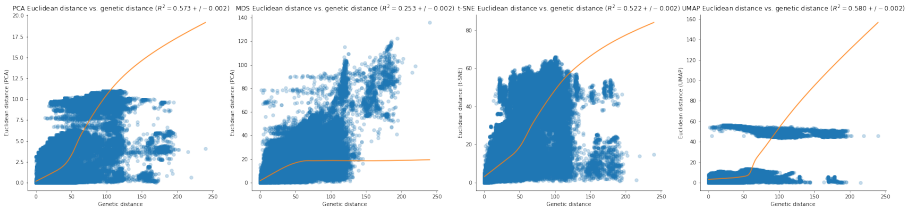
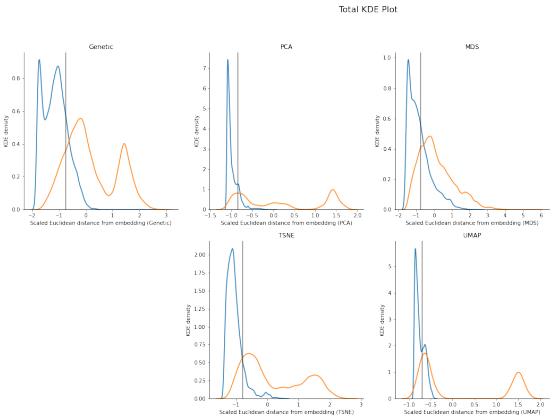


Figure Six



MERS:

MERS: MERS in this project is used as a extreme test case. MERS is an incredibly recombinant virus, making it difficult to construct a phylogenetic tree explaining the relationships between strains and locations. The evolutionary rate of MERS is within the expected range for RNA viruses, with a rate from Baysian structured coalescent estimate of around 9.57 * 10ˆ-4 (95% Highest Posterior Density: 8.28 - 10.9 * 10ˆ-4) subs/site/year. Observed departures from strictly clonal evolution suggest 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. The genomes are 30130 bases long, with a

mean bases missing of 889.781 and median of 42.5. It can be reasonably assumed that MERS is a good case to test and challenge Cartography.

SUMMARY OF RESULTS FOR MERS:

SUMMARY OF RESULTS ACROSS VIRUSES

Overall, the best recapitulation of the phylogenetic clades of the four analyzed was that of MDS and UMAP, because they preserved the most local and global structure. However, t-SNE separated clusters much better than MDS as lots of the points were layered on top of each other in the MDS embedding (clade A1b/131K in h3n2 Influenza was impossible to see in MDS clusters 1 and 2).

Of the 3 embeddings that reduced the Hamming distance matrix, as the focus of the algorithms shifted more towards preserving local structure over global structure, the closer the Pearson Coefficient got to 0 (MDS > UMAP > t-SNE). Pearson Coefficient studies the effectiveness of an embedding at preserving a relationship between genetic and euclidean distance, so for t-SNE, an algorithm that focuses primarily on exaggerating distances and clusters locally to convey patterns, the pearson coefficient is going to be closer to 0, as the data points will not adhere to a best fit line. For MDS, however, the embedding relies almost entirely on creating an exact 1:1 genetic:euclidean relationship, so the pearson coefficient was much higher. In the same vein, as the algorithms shifted towards retaining local patterns over global patterns, the disparity between the densities of the between vs within violin plots became more pronounced. Because the violin plots assess an embedding's ability to distinguish between clades (how clustered the embedding is), the more exaggerated the differences between euclidean and genetic distance, the more disparate the densities are. t-SNE's within clade violin plot had the most concentrated density at around 5, and its between clade violin plot had the most concentrated density at around 45. By comparison, the genetic distance within:between was 45:60.

Discussion

Supplementary Figures and Analysis

Explained Variance Plots for PCA

Flu

Zika

MERS

PCA Full Plots

Flu

Zika

MERS

MDS Full Plot:

Flu

Zika

MERS

Works Cited

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

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. Public Library of Science. 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. eLife Sciences Publications, Ltd. https://elifesciences.org/articles/31257.

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.

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*. Nature. https://doi.org/10.1038/nature22402.

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

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

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*. U.S. National Library of Medicine.

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.)*. U.S. National Library of Medicine.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction."

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*. U.S. National Library of Medicine.

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*. Elsevier.

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). National Academy of Sciences: 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. Nature. https://www.nature.com/articles/nature06945.