AN INTEGRATED PHYLOGEOGRAPHIC ASSESSMENT OF THE BANTU MIGRATION USING MACHINE LEARNING AND STATISTICAL VALIDATION

by

Colby Tyler Ford

A dissertation proposal submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2017

Approved by:

_____
Dr. Daniel Janies

_____
Dr. Xinghua Shi

_____
Dr. Anthony Fodor

_____
Dr. Mirsad Hadzikadic

_____
Dr. Matthew Parrow

ABSTRACT

COLBY TYLER FORD. An Integrated Phylogeographic Assessment of the Bantu Migration using Machine Learning and Statistical Validation. (Under the direction of DR. DANIEL JANIES)

The Bantu grould is used to describe a category of peoples for around 600 different ethnic groups in Africa ranging from Cameroon down to South Africa. Research in the 1970's was completed to geographically divide the Bantu languages into sixteen zones now known as "Guthrie Zones" [16]. Recently, researchers have postulated the migratory pattern of the Bantu people by looking at cultural information or some genetic factors. This has proven to be an area of disagreement as the resulting phylogenies can differ drastically, depending of the type of data taken into account.

In this proposal, I am proposing to make my own assessment of the Bantu migration using cultural data combined with genomic data. My hypothesis is that Bantu expansion can be characterized by a primary split in lineages, which occurred early on and prior to the population spreading south through the now Congolese forest region. The primary goal is to generate a comprehensive phylogenetic tree as a model of Bantu migration using Y-chromosomal, mitochondrial, and cultural data. This, together, should result in a more robust, multifaceted migratory model. This differs from the current methods as they each only use a single data type, which limits the reach of the analysis and introduces additional bias into the mix. However, by integrating different data together, the migration pattern can now be drawn from different factors at once.

For validation purposes, I am proposing multiple computational, machine learning,

and statistical methods. Unsupervised machine learning methods such as clustering and multidimensional scaling will be used to test the strength of migratory model. Mainly, these feature extraction methods will test the agreement between the model as well as the machine learning results. Results from the machine learning steps will be shared visually for review and comparison. The visual comparison of this migratory model to the current models will be performed by mapping the models geographically. Computational and statistical methods such as molecular sequence alignment, heuristic tree space search, plus bootstrapping and cross-validation will be used to further assist in the migratory model generation and testing process.

My resulting migratory model will likely be discordant to the findings of at least one of the other phylogenies that use a singular type of data, but will be supported by means of machine learning and statistical methods. Depending on the validation results, this should suggest that the combined model is likely closer to the true historical Bantu migration. Thus, the results of the analysis will allow for a ruling on my hypothesis around the location of the initial Bantu split.

This dissertation expands upon work that I have performed while researching with Dr. Daniel Janies and other researchers from the American Museum of Natural History and The University of North Carolina at Charlotte.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND BACKGROUND

On the continent of Africa, approximately one-third of the population falls under the category of Bantu. Bantu is a group of over 150 million people from central and southern Africa. Among the Bantu population, there are around 600 languages (including dialects) spoken. It is known that the Bantu people originated from what is now Cameroon and likely spread to the east and south over time [7]. However, the exact migratory or expansion path that was taken is unknown and, as a result, the point of some debate.

By looking at changes in the Bantu people geographically, we can approximate the trajectory at which the individuals moved around continent throughout time. But, depending on the type of changes that are in question, the resulting migratory path can differ drastically.

1.1 Current Migratory Models

Recently, researchers from various disciplines have published their own postulations around the Bantu migration. While there is overwhelming agreement that the group started in what is now Cameroon, the agreement stops there. When comparing the publications on this topic, it is easy to notice that the migration paths are very different from one another. This is likely due to data that was used by each research team. Some researchers have relied on simplified genetic data (such as single nucleotide

polymorphisms) where as others have taken a more culture-driven approach. Some of the cultural approaches include using linguistic data or even data around cultural behaviors such as marriage practices or dependencies on hunting and gathering, for example. While these datasets and approaches are valid, they paint very different pictures about the migration path.

Linguistic researchers such as Dr. Rebecca Grollemund and others have shown that there is considerable importance in linguistic traits of Bantu languages as markers for inferring migratory patterns [15]. Other research has surfaced that suggests that the hypothesis concerning Bantu expansion can be tested by using both linguistic and some genetic data [10]. As shown on the maps in figure 1, there is some obvious disagreement in the proposed Bantu expansions. There has even been some research that attempts to uncover a Bantu expansion by analyzing the spread of farming [27] across the continent.

1.2 Hypothesis on the Early Bantu Split

My hypothesis is that the Bantu expansion can be characterized by a primary split in lineages, which occurred early on and prior to the population spreading south through the now Congolese forest region. In maps (a) and (b) in figure 1, the de Filippo models show two different points of divergence, one beginning in present-day Cameroon and one with a later split, following a more linear migration path. If the split was indeed earlier on, then the divergence and migration should flow from north to south. That is, the Bantu individuals in present-day South Africa are the most dissimilar (both culturally and genetically) to the Bantu individuals in present-day

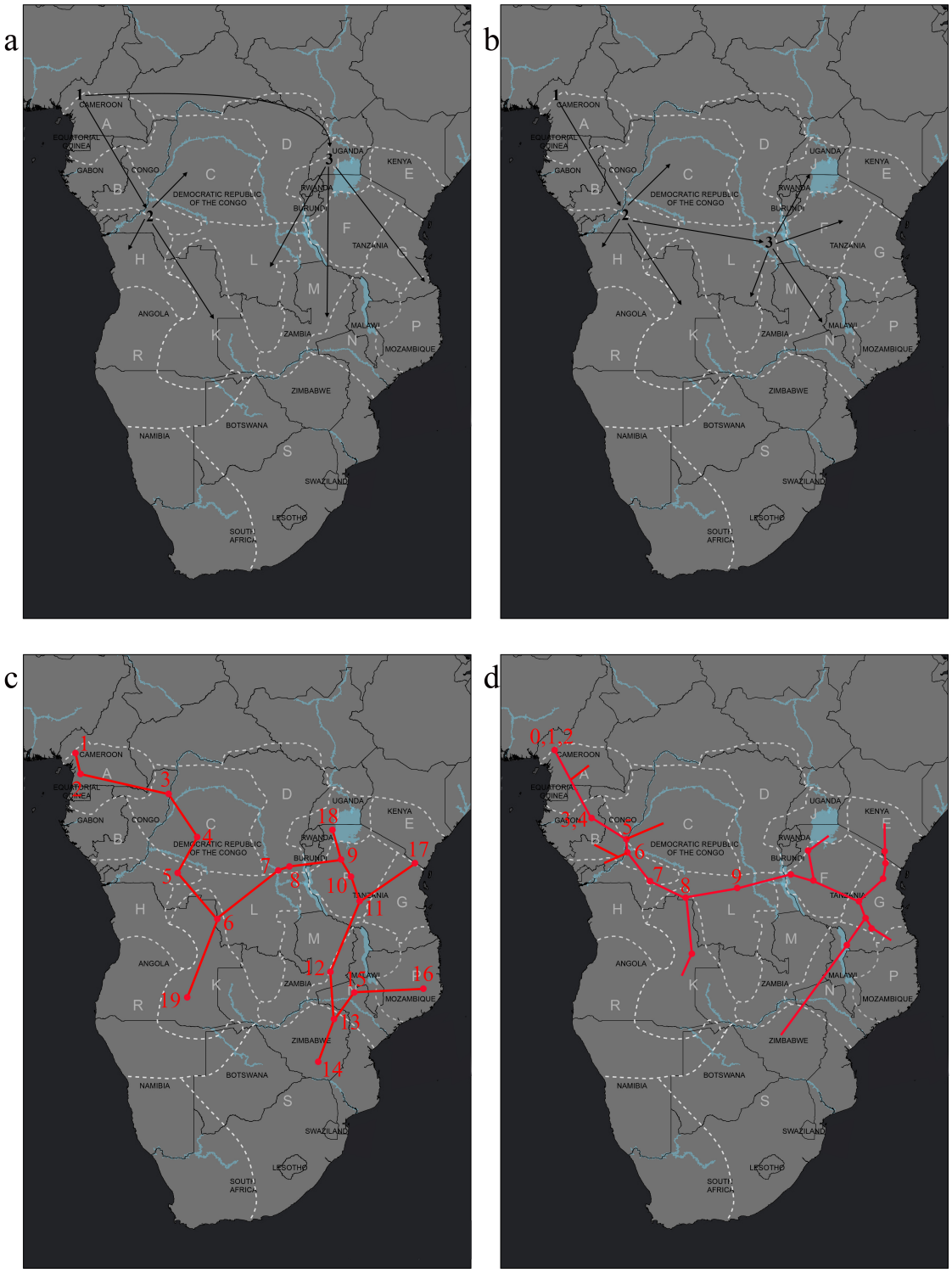

Figure 1: Hypotheses of Bantu language expansion
a) early split vs. b) late split de Filippo et al., 2011 [10], (fig. 2)
c) Currie et al., 2013 [9], (fig. 2b)
d) Grollemund et al., 2015 [15], (fig. 2A) main nodes and branches.

Cameroon, correlated to their geographic distance.

By combining the cultural and genomic data to create a single migratory path, this hypothesis can be tested. Also, it will be of interest to see the concordance or discordance from this model to the other migratory models that are currently available.

1.2.1 The Congolese Forest as a Geographic Barrier

A sub-hypothesis around the migratory path is that the Congolese forest has provided a geographic boundary of sorts, around which the Bantu peoples migrated. As shown in figure 2, the rainforest is shown to run through southern Cameroon, Gabon, Republic of Congo, and the Democratic Republic of Congo.

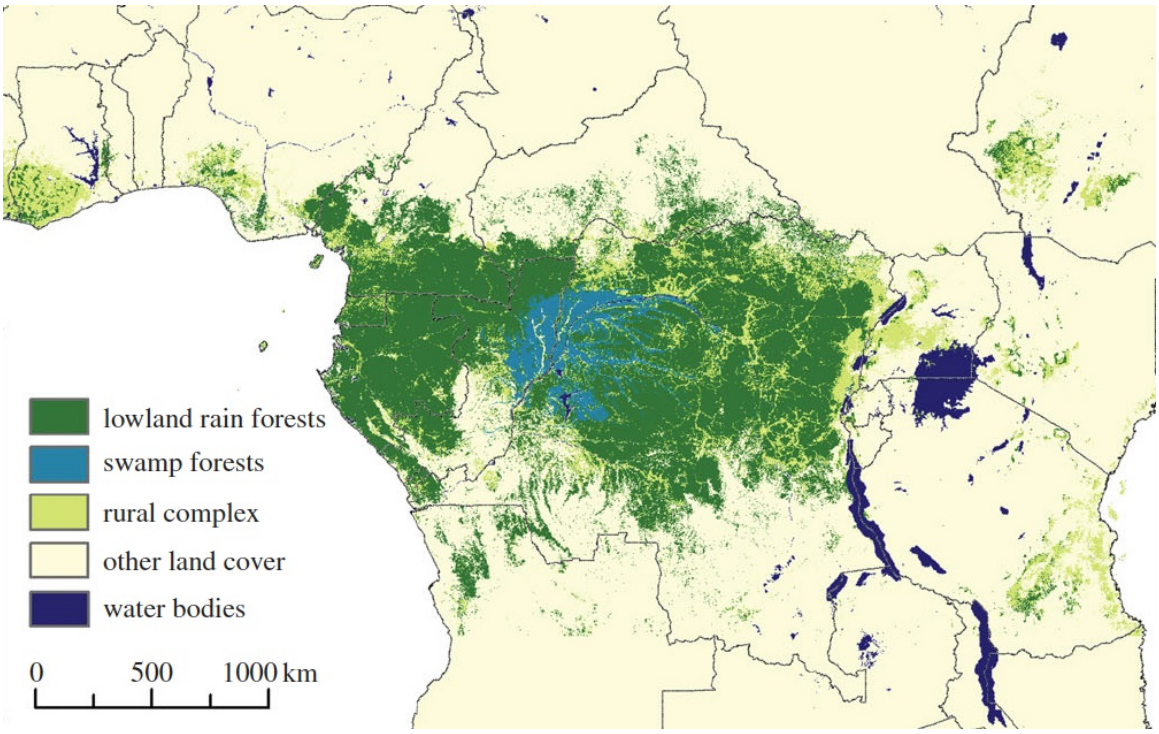

Figure 2: Spatial distribution of the African rainforests derived from MODIS data [6].

If my initial hypothesis is correct and the Bantu expansion had an early split that

went around the Congo area, further work could be done to understand if this was truly due to the rainforests prohibiting migrant travel and settling or if other factors played a role.

CHAPTER 2: MIGRATORY MODEL GENERATION

2.1 Introduction

To begin analyzing the migratory path of the Bantu population, we can rely on phylogenetic analyses to create a model tree. However, certain data processing exercises must me completed before the data are ready to use in tree generation. The required processing steps may differ depending on the input data that are available.

2.2 Data Curation

Genetic data has been curated for Y-chromosomal short tandem repeats (STRs) and mitochondrial DNA (mtDNA) from various papers on the subject matter. As of today, there are 79 mtDNA samples and 59 Y-chromosomal STR samples. These samples cover 56 and 49 distinct language groups (plus 2 outgroups), respectively. For some previous Bantu migration publications, I noticed that there is an uneven distribution of the samples among the different groups, which may have introduced bias into their respective models. For example, in the de Filippo migration study [10], figure 2 in the publication shows the heavy concentration of data samples in the northeastern area around Cameroon. To check for a geographic bias of this data, the sampling locations from each of the papers were mapped. See figure 3. The spread of my genetic samples prove to be more evenly distributed compared to other publications' datasets.

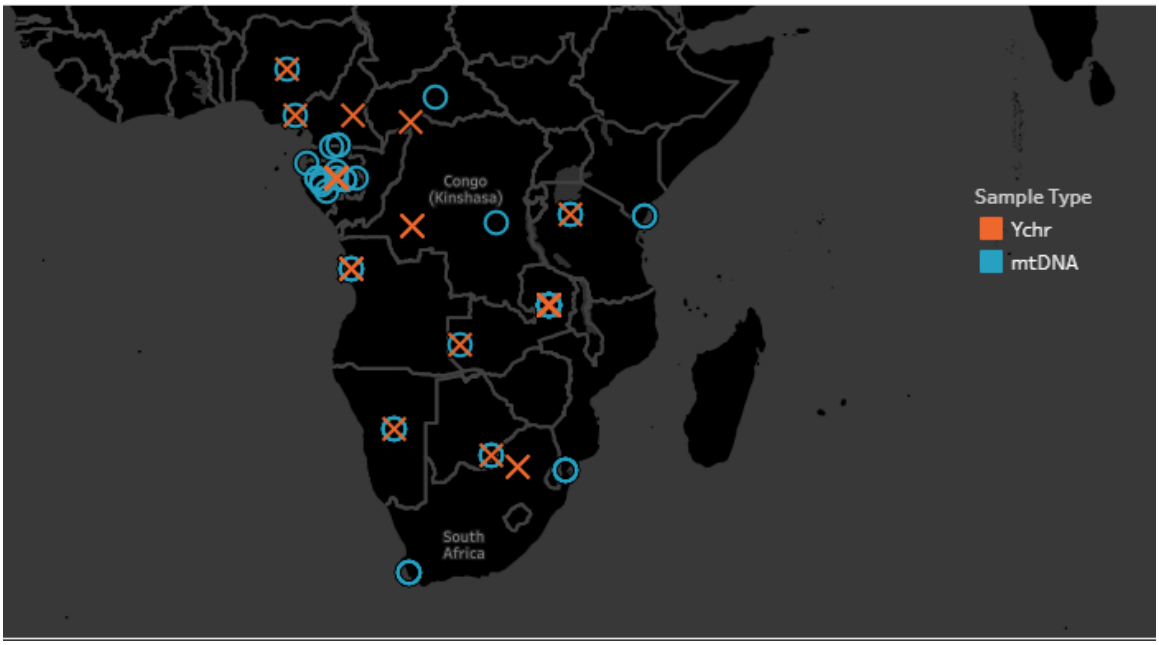

Figure 3: Mapped locations of the 138 genetic samples.

In addition to the genetic data, linguistic and cultural data have been collected and will be used in conjunction with the mtDNA and STR data. For the linguistic data, I have 91 language groups represented. For cultural, 93 are represented. Both datasets have 15 samples that are treated as outgroups.

The linguistic data consists of frequency metrics of 255 different sounds. This data is based on the Swadesh-100 list of words, producing a lexical dataset, which was rendered into LaTeX TIPA format. The cultural data is comprised of 92 cultural traits outlined by the Ethnographic Atlas Codebook [14]. These "traits" are cultural customs such as the dependence on hunting, gathering, and fishing, as well as marriage practices, the use of animals for food and agriculture, and gender roles.

Among the four datasets, only 20 of the groups are an exact match in all of them. However, despite having a low number of specific groups matching, almost all of the Guthrie zones represented in each dataset. Zones D,E,F,G,J, and P are underrepre-

sented in the genetic data only. All zones are represented in the linguistic and cultural datasets. See table 1. As a result, I will only postulate a migratory model at the Guthrie zone level using aggregated data from groups withing the zones.

For a complete list of citations for the data, see table 2.

Guthrie Zone	mtDNA	Ychr	Cultural	Linguistic	Total
A	4	3	5	5	**17**
B	8	13	5	5	**31**
C	2	2	11	11	**26**
D	0	0	5	5	**10**
E	2	0	6	7	**15**
F	0	1	3	3	**7**
G	2	0	7	7	**16**
H	1	2	6	6	**15**
J	0	0	9	9	**18**
K	7	7	3	4	**21**
L	3	3	2	2	**10**
M	6	5	6	6	**23**
N	4	3	6	6	**19**
P	0	0	2	2	**4**
R	3	3	3	3	**12**
S	14	7	12	12	**45**
Z (Outgroup)	2	2	12	12	**28**
Total	**58**	**51**	**103**	**105**	**317**
Zone Coverage	*75%*	*69%*	*100%*	*100%*	*100%*

Table 1: Representation of Guthrie zones among datasets.

Sample Type	Citation
Mitochondrial	Barbieri, Chiara, et al. "Migration and interaction in a contact zone: mtDNA variation among Bantuspeakers in southern Africa." PloS one 9.6 (2014): e99117.
Mitochondrial	Behar, Doron M., et al. "The dawn of human matrilineal diversity." The American Journal of Human Genetics 82.5 (2008): 1130-1140.
Mitochondrial	Castr, Loredana, et al. "mtDNA variability in two Bantuspeaking populations (Shona and Hutu) from Eastern Africa: Implications for peopling and migration patterns in sub-Saharan Africa." American journal of physical anthropology 140.2 (2009): 302-311.
Mitochondrial	de Filippo, Cesare, et al. "Genetic perspectives on foragerfarmer interaction in the Luangwa Valley of Zambia." American journal of physical anthropology 141.3 (2010): 382-394.
Mitochondrial	Gonder, M. K., Mortensen, H. M., Reed, F. a, de Sousa, A. & Tishkoff, S. A. Whole-mtDNA genome sequence analysis of ancient African lineages. Molecular Biology and Evolution 24, 757768 (2007).
Mitochondrial	Gonder, M. K., Mortensen, H. M., Reed, F. a, de Sousa, A. & Tishkoff, S. a. Whole-mtDNA genome sequence analysis of ancient African lineages. Molecular Biology and Evolution 24, 757768 (2007)., 9(Mary2005 in MtDNA_JIA): 9. Mishmar, D. et al. Natural selection shaped regional mtDNA variation in humans. Proceedings of the National Academy of Sciences of the United States of America 100, 171176 (2003). Rito, T. et al. The first modern human dispersals across Africa. PloS one 8, e80031 (2013).
Mitochondrial	Horai, S. & Hayasaka, K. (1990)Am. J. Hum. Genet. 46,828-842. Brucato, Nicolas, et al. "The imprint of the Slave Trade in an African American population: mitochondrial DNA, Y chromosome and HTLV-1 analysis in the Noir Marron of French Guiana." BMC evolutionary biology 10.1 (2010): 1.
Mitochondrial	Mishmar, D. et al. Natural selection shaped regional mtDNA variation in humans. Proceedings of the National Academy of Sciences of the United States of America 100, 171176 (2003).
Mitochondrial	Quintana-Murci, Lluis, et al. "Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy huntergatherers and Bantuspeaking farmers." Proceedings of the National Academy of Sciences 105.5 (2008): 1596-1601.; Behar et al (2008)
Y-chromosomal	Allele frequencies and haplotypes for 28 Y-STRs in Ovambo population. Legal Medicine
Y-chromosomal	Berniell-Lee, Gemma, et al. "Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages." Molecular biology and evolution 26.7 (2009): 1581-1589.
Y-chromosomal	Coelho, Margarida, et al. "On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola." BMC evolutionary biology 9.1 (2009): 1.
Y-chromosomal	de Filippo, Cesare, et al. "Genetic perspectives on foragerfarmer interaction in the Luangwa Valley of Zambia." American journal of physical anthropology 141.3 (2010): 382-394.
Y-chromosomal	de Filippo, Cesare, et al. "Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups." Molecular biology and evolution 28.3 (2011): 1255-1269.
Y-chromosomal	Henn, Brenna M., et al. "Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa." Proceedings of the National Academy of Sciences 105.31 (2008): 10693-10698.
Y-chromosomal	Leat, Neil, et al. "Properties of novel and widely studied Y-STR loci in three South African populations." Forensic science international 168.2 (2007):154-161?
Y-chromosomal	Lecerf, Maxime, et al. "Allele frequencies and haplotypes of eight Y-short tandem repeats in Bantu population living in Central Africa." Forensic science international 171.2 (2007): 212-215.
Y-chromosomal	Tishkoff, Sarah A., et al. "History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation." Molecular biology and evolution 24.10 (2007): 2180-2195.
Cultural	Gray, J. P., "Ethnographic Atlas Codebook" World Cultures 10.1 (1998): 86-136.
Linguistic	Bastin, Y., Coupez, A., and Mann, M. (1999). Continuity and Divergence in the Bantu Languages: Perspectives from a Lexicostatistic Study. Annales du Muse Royal de l'Afrique Centrale, Sciences Humaines, #162. Tervuren.
Linguistic	Whiteley, P., Wheeler, W., and Xue, M. (2017). Revising the Bantu tree: Words as sequences 2.0. Current Anthropology. in review.

Table 2: List of data sources.

2.3 Data Workflow

Given the diversity of the data, each dataset must be standardized such that each can be similarly analyzed. Also, to combine the datasets for a combined analysis, each dataset must be converted so that the joining can occur. The optimal format for both standalone and combined analyses is the .nexus file format. Nexus is an extensible file format that is quite popular in the bioinformatics field as it is flexible enough to house different types of data and metadata [23].

The STR data was derived by hand from the original Y-chromosomal samples and manually formatted as a .nexus file. For the mitochondrial data, the original file format was an unaligned .fasta file. Multiple sequence alignment was performed using MAFFT [18] (with default settings) and the results were exported as an aligned .nexus file. Some manual cleansing of the mitochondrial data was necessary as there were some sequence quality issues in the original .fasta file(s). The cultural data was attained pre-formatted as a .nexus file. No changes were made here.

The linguistic data was generated via a separate research project. The linguistic dataset comes from sub-Saharan Bantu (and Bantoid outgroup) languages from sub-Saharan Africa collated at the Royal Museum for Central Africa [4], extended by Whiteley et al. [36]. IPA renderings (in LaTeX TIPA format) of each word were created and treated as sound sequences. See figure 4.

See figure 5 for a graphical representation of the workflow.

Once the data conversions are complete, the datasets are run through at least one a several tree generation tools. The set of tools that may be used includes PAUP

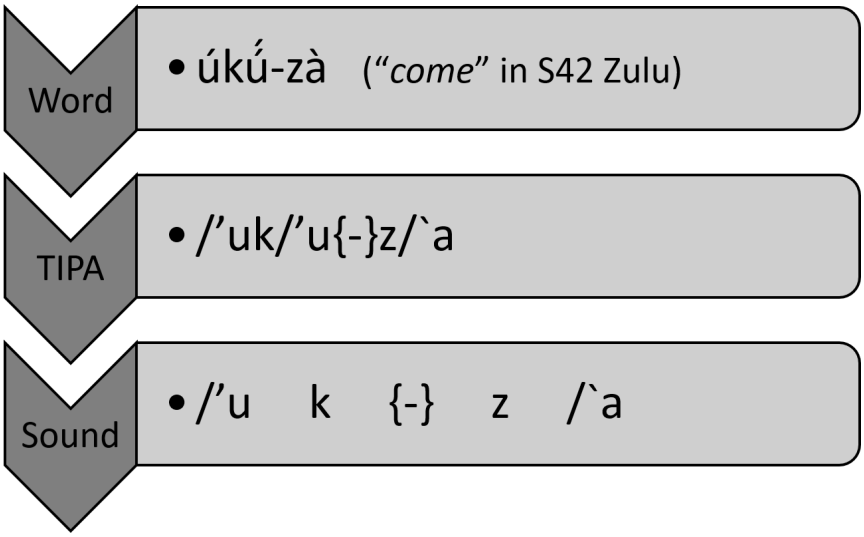

Figure 4: Production of the LaTeX TIPA forms of Words

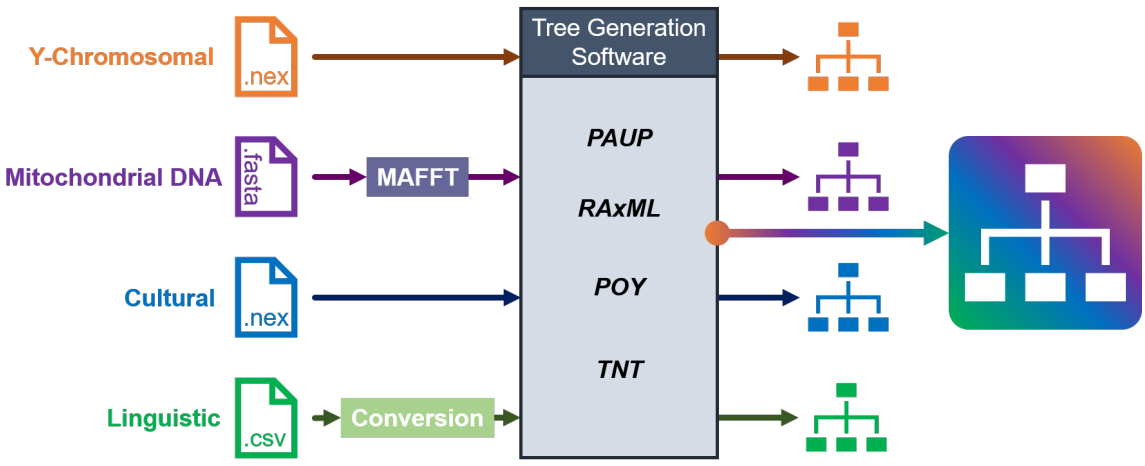

Figure 5: Tree Generation Workflow.

[32], RAxML [31], POY5 [35], and TNT [12]. PAUP is likely the tool of choice for the combined tree as it is flexible enough to handle non-genomic data along with the STR and mtDNA data. Plus, PAUP has an extensive heuristic tree space search. This will allow for a single, heuristically-derived consensus tree to be created using the parsimonios methd and all available data.

CHAPTER 3: ANALYSIS METHODS

3.1 Introduction

In addition to phylogenetic tree generation, I am proposing to employ multiple machine learning and statistical methods to analyze the data. Phylogenetic trees use inference methods that rely on an optimality criterion such as maximum likelihood or maximum parsimony. Along with these inference methods, I will use distance-based methods to create trees as well.

Distance-based tree generation methods such as UPGMA [30] or the Neighbor-Joining method [28] use distance or similarity matrices to recursively combine objects into a tree. In this case, they would compare distances/similarity for each taxon, then continually combine them to form a single tree. Distance-based methods are very flexible in that the distance input can be from virtually any source. Thus, combining a distance-based method with resulting data from the below machine learning or statistical methods should prove effective at generating a robust set of additional trees for comparison to the other inference-based methods.

Each of the methods below process the data in a slightly different way. So, agreement between the outputs of the tree generation and each of the other methods will build support or confidence in my Bantu migration model. If there are drastic differences between the result, it should then help to pinpoint the areas where confidence

is low in the migration model.

3.2 Feature Reduction and Clustering Methods

3.2.1 Multidimensional Scaling

Multidimensional Scaling (MDS) is a technique usually used to reduce dimensionality [20]. MDS can accept an input of a correlation matrix, distance matrix, or a similarity matrix. Then, the number of dimensions for the analysis must be selected *a priori*. I will run the analysis using a sweep of dimensions from 2 to $n-1$, where n is the number of dimensions in the input data. However, only MDS results in 2 or 3 dimensions can be visualized.

The steps of the classical MDS algorithm are as follows:

1. Begin by calculating the Eucliean distance between two points i and j to form the distance matrix D.
2. Using D, calculate $A = \left\{-\frac{1}{2}{d_{ij}}^2\right\}$.
3. Then, calculate $B = \{a_{ij}a_{i.}a_{.j} + a_{..}\}$, where $a_{i.}$ is the average of all a_{ij} across j.
4. Find the p largest eigenvalues $\lambda_1 > \lambda_2 > ... > \lambda_p$ of B plus their corresponding eigenvectors $L = \{L_1, L_2, ..., L_p\}$.
5. Output the coordinates of the objects, which are the rows of L.

Then, The goal is to place the objects in an N-dimensional space such that the distances between each object are maintained as well as possible. To do this, MDS minimizes a cost function known as "stress", shown in equation 1, which is a residual sum of squares metric.

$$stress = \sqrt{\frac{(d_{ij} - \hat{d}_{ij})^2}{\sum \hat{d}_{ij}^2}} \quad (1)$$

If the $stress \leq 0.05$, the goodness-of-fit is considered acceptable. However, the number of dimensions that yields the lowest stress will be considered the best number of dimensions for the analysis of the data. The result of an MDS analysis is an multidimensional projection of the data that has measured the similarity of the data points. In this case, a matrix where each taxon (on rows) is described by a number of dimensions (on columns).

Classical MDS is available as an R function in the *stats* packages called *cmdscale* [25].

3.2.1.1 Parallel Trajectories of Genetic and Linguistic Admixture in a Genetically Admixed Creole Population

Research performed by Verdu et al. [34] around the genetic and linguistic trajectories of Creole people will serve as a model for my analysis of the Bantu data. The researchers used multidimensional scaling (MDS) analysis to look at individual-pairwise allele-sharing dissimilarities of different populations. Also, the researchers used the k-Means clustering method to analyze admixture among the groups. From a visualization standpoint, the graphics that are generated will be quite useful in my overall assessment of variation. See figures 6, 7, and 8. I plan to replicate these visuals using the Bantu data.

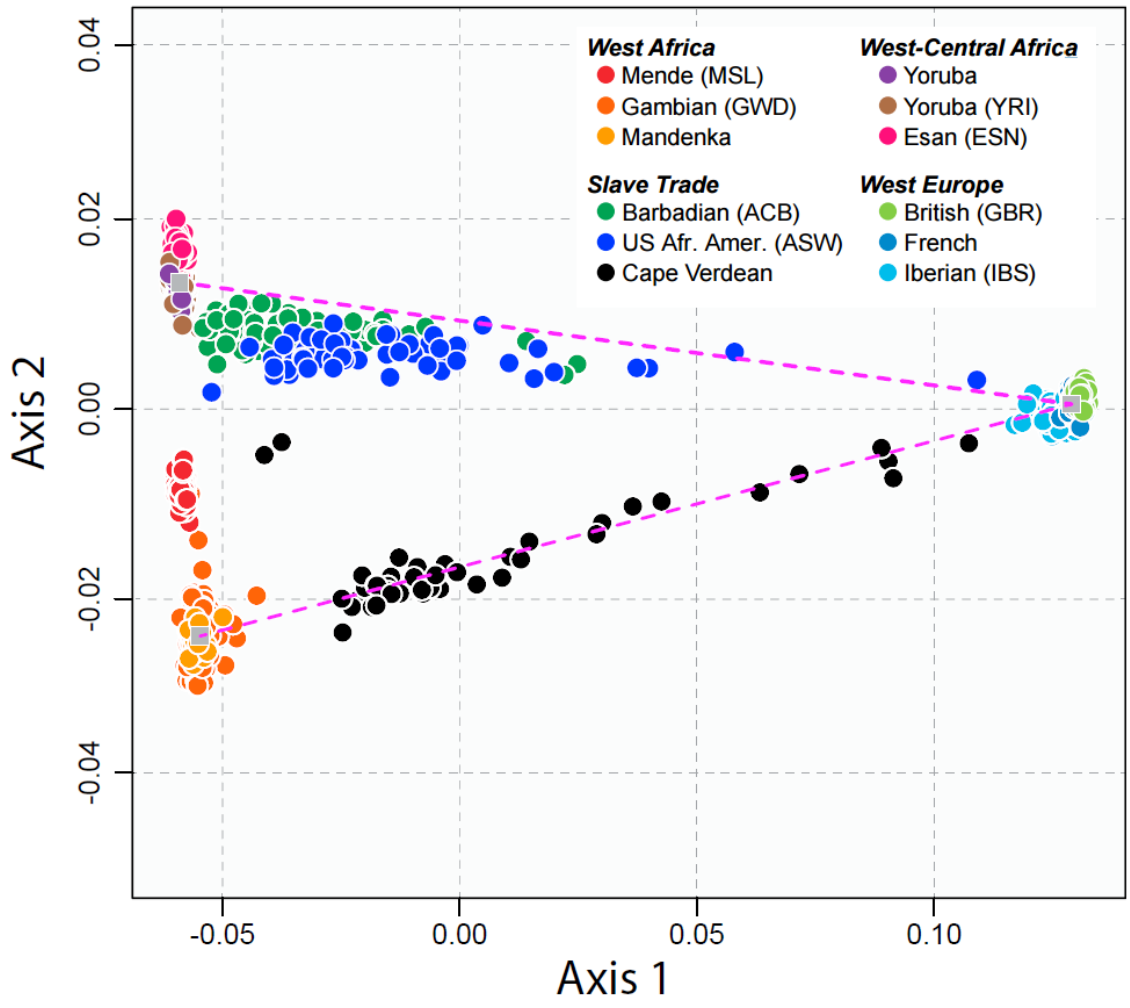

Figure 6: MDS analysis of six continental African populations from West and West-Central Africa, three Western European populations, and three admixed populations that arose during the transatlantic slave trade. ([34], figure 1A)

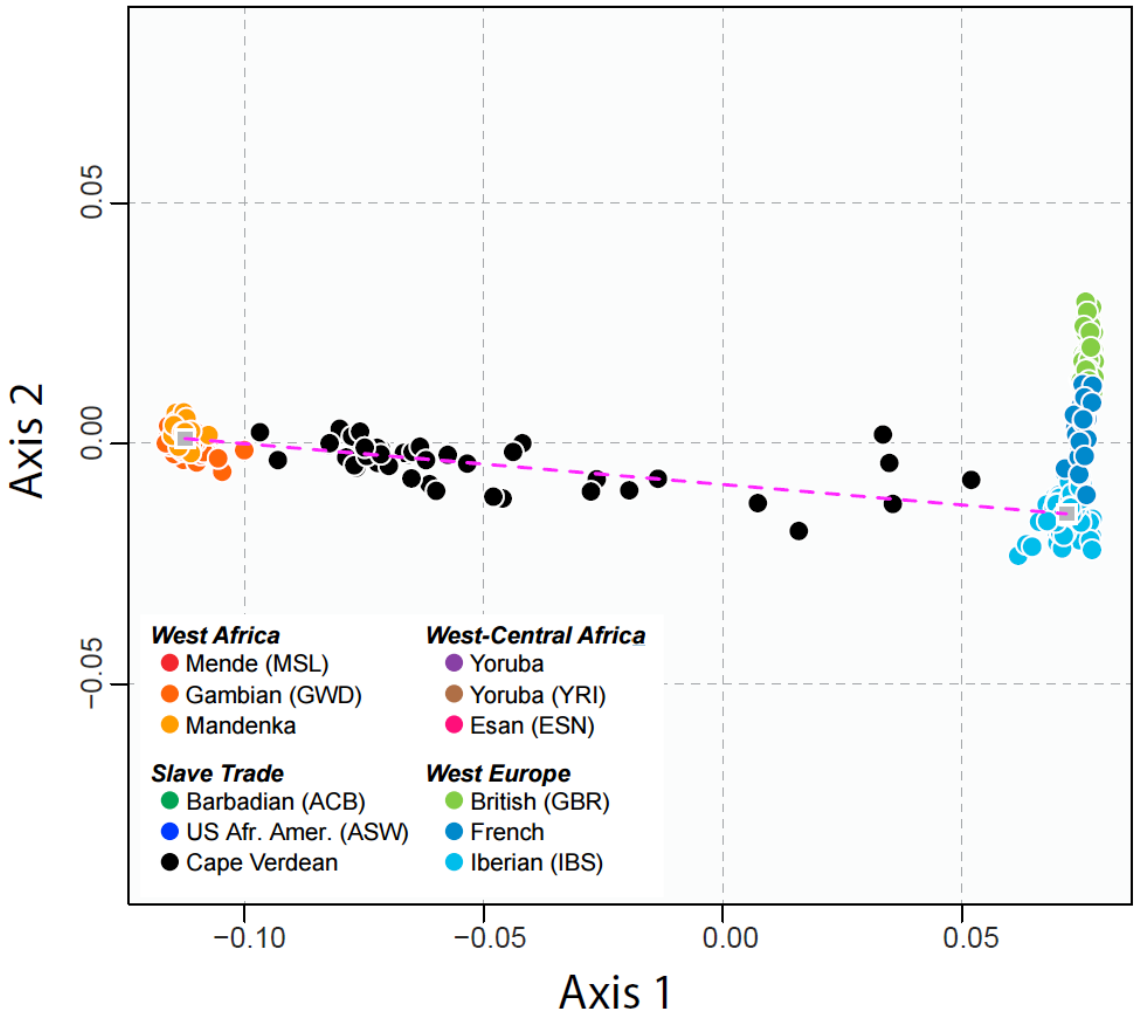

Figure 7: MDS analysis of two West African populations (Gambian Mandinka and Senegalese Mandenka), three European populations, and the Cape Verdean population. ([34], figure 1B)

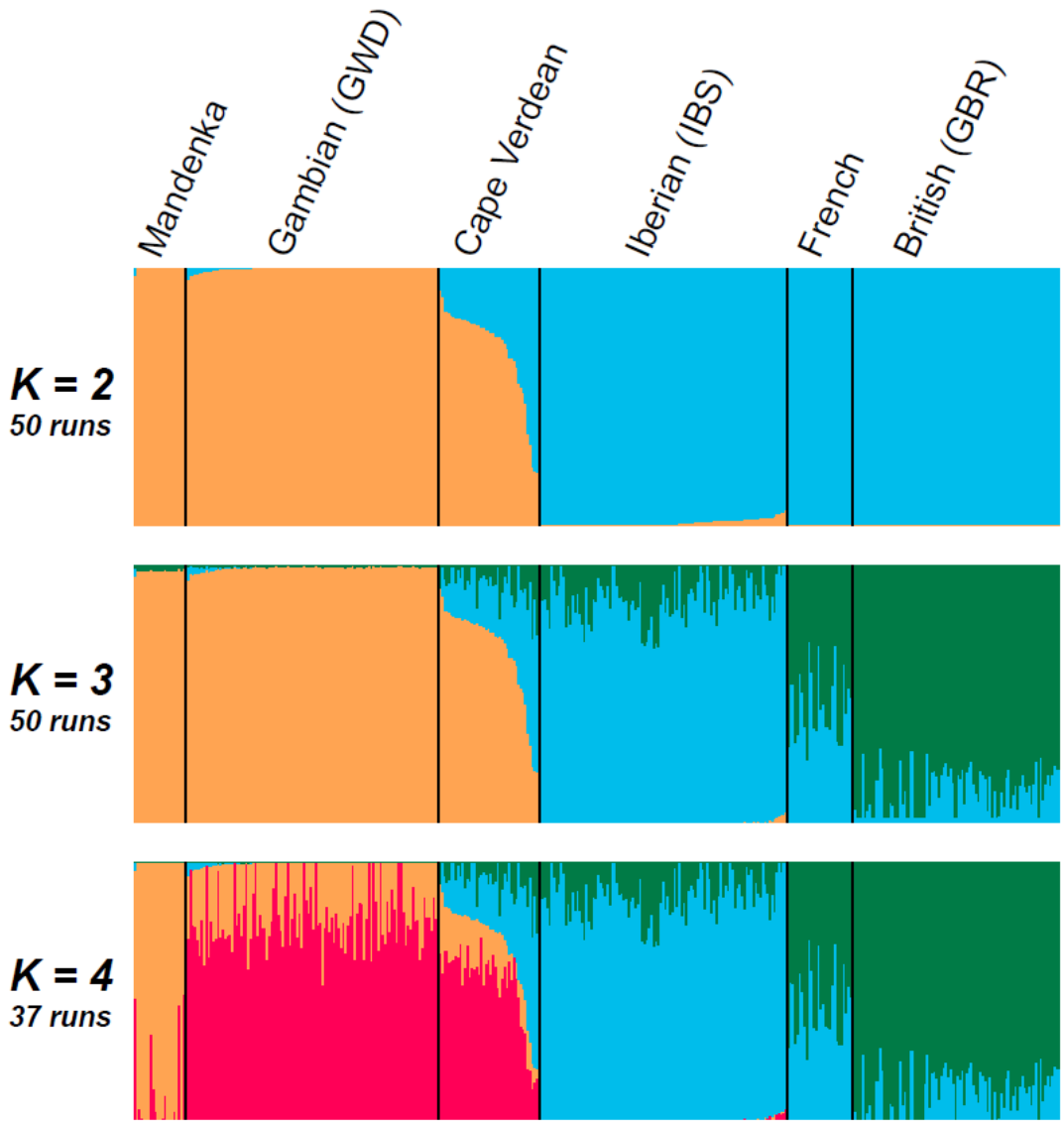

Figure 8: Admixture analysis of Cape Verdean individuals together with West African and Western European samples. ([34], figure 1C)

3.2.2 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is an algorithm designed to reduce dimensionality while taking the correlations between two sets of variables into account [26]. Traditionally, CCA is performed on only two variable set. However, it is possible to generalize CCA for more than two variable sets using a probabilistic interpretation [3].

CCA is interested in maximizing the correlation of multiple variable sets. This is completed by computing cross-covariance matrices and their corresponding projection vectors. Then, selecting the highest eigenvectors with the highest eigenvalues.

CCA is quite related to Principle Component Analysis [11] in some of the matrix calculations as well as the final step of choosing the necessary components. To decide how many pairs of eigenvectors to use, we look at the relative value of the corresponding eigenvalue for each. Then, keep enough of the eigenvectors to conserve the correlation in the data sufficiently. The new, lower-dimensional representation of the data can now be used from the new vector of (a_i, b_i) pairs. See figure 10.

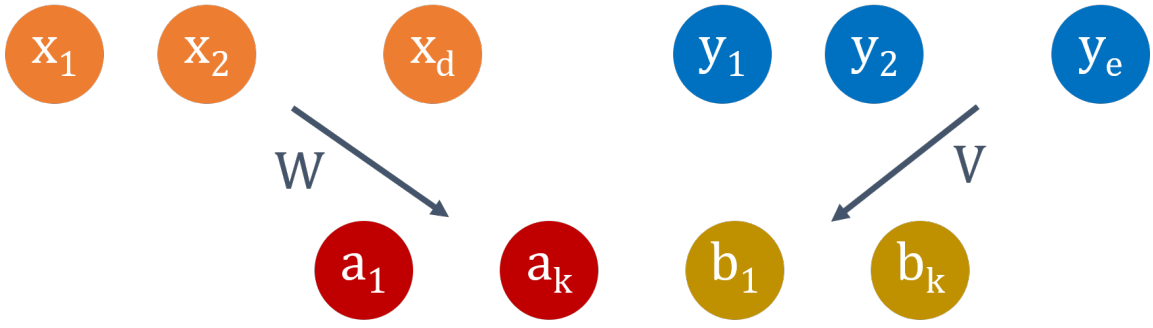

Figure 9: Canonical correlation analysis example using two sets of variables x and y. Showing that the correlation is maximized after projection.

CCA will be used to help understand the covariation of the traits from different

datasets. For example, are there STRs in the Y-chromosomal data that are heavily correlated with a marriage practice in the cultural data. Heavy amounts of correlation in different traits for certain group may help to understand the migration trajectory that said groups took.

CCA is available as an R package called *CCA* [13].

3.2.3 Laplacian Eigenmaps

As a minor sub-comparison to MDS and, by extension, the trees, I plan to use a Laplacian Eigenmap (LE) [5]. While MDS may provide an acceptable projection of the data in a lower-dimensional space, but an LE may provide a better visual of the similarities in the taxa. Traditionally, LE's provide a denser placement of similarly-related data points and a sparser placement of distantly-related cases. Note the visual differences between MDS and LE on the iris sample dataset in figure 10, for example.

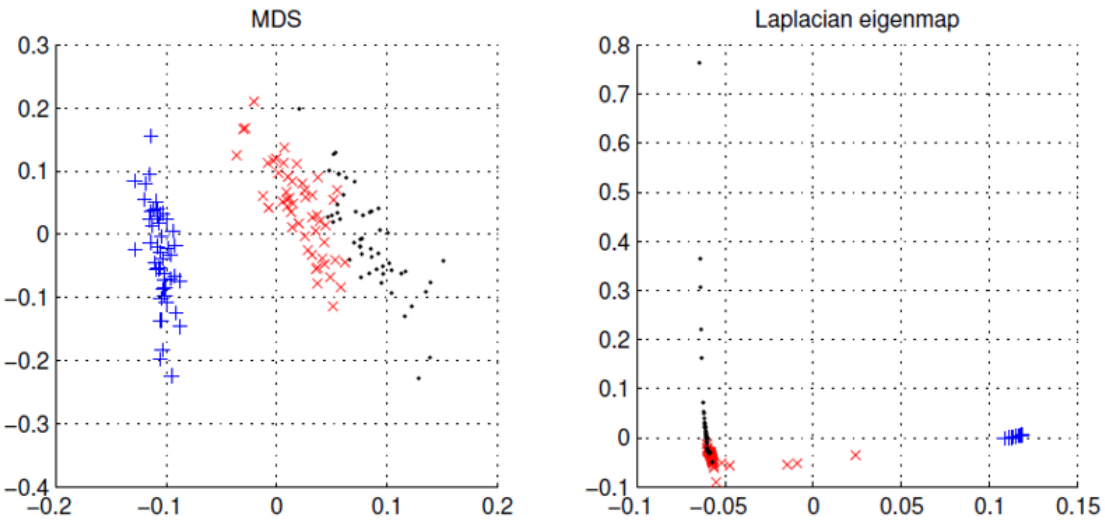

Figure 10: Iris data reduced to two dimensions. Method: Left - Multidimensional Scaling, Right - Laplacian Eigenmaps. The Laplacian eigenmap lends a denser placement of similar iris flowers [2].

LE's look for local similarities only. This is accomplished by defining a neighborhood using the k-Nearest Neighbor algorithm, the epsilon neighborhood algorithm, or some other distance matrix method. Using the Gaussian kernel, Euclidean distances are converted to similarity values. The desired number of reduced dimensions must be defined *a priori*. I will run the analysis using 2 and 3 dimensions given that this is the limit to what can be visualized.

LE is implemented as an R package called *dimRed* [19].

3.2.4 k-Means Clustering

As a final machine learning method, I plan to use k-Means clustering [21], [22]. k-Means is an unsupervised machine learning algorithm that heuristically (the problem is NP-hard) clusters variables into a set number of groups solely based on their mathematical distance from each other. Traditionally, this method is used to discover distinct groups in a data set. However, this clustering method can also be used for dimensionality reduction a well. The groups are represented by their centroid locations and can then be compared to see which groups are closer other other groups, thereby implying a relationship or similarity between them.

Being an unsupervised algorithm makes K-means clustering quite flexible to many types of data. It is agnostic of the physical nuances of the columns and simply treats each factor as a dimension. Then, it uses the Euclidean distance between the points in the algorithm. Given that the Bantu data will include many different types of data from different sources, the k-Means clustering algorithm is an obvious choice to uncover groups in the data.

The k parameter (for number of clusters) must be defined *a priori*. For visualization purposes only $k\epsilon\{2,3\}$ can be used. However, the optimal number of clusters may be greater than 3. To determine the optimal number of clusters, one of two methods will be used: the *elbow method* and the *gap statistic* [33].

For the *elbow method*, the sum square of error (SSE) will be calculated for each level of k (from 2 to $n-1$, where n is the number of dimensions in the input data). The SSE is then plotted as a line in relation to the k levels. The location of a bend (the elbow) in the plot is the indicator for the the appropriate value of k, the number of clusters.

For the *gap statistic*, the steps are as follows:

1. Cluster the dataset for each value of $k = 2, ..., n-1$ and compute the corresponding total within intra-cluster variation W_k.

2. Generate B reference sets, each with a random uniform distribution. Cluster each of these reference sets with varying number of clusters $k = 2, .., n-1$ and compute the corresponding total within intra-cluster variation W_{kb}.

3. Compute the estimated gap statistic as the deviation of the observed value W_k from its expected value W_{kb} or $Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}^*) - log(W_k)$. Also, compute the standard deviation of the statistics.

4. Choose the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at $k+1$: $Gap(k) \geq Gap(k+1)s_k + 1$.

k-Means clustering is available as an R function in the *stats* packages called *kmeans* [25].

3.2.5 Hierarchical Clustering

My goal is to use each of these methods to compare to the phylogenetic analysis and, therefore, the resulting trees. All of the aforementioned methods result in a matrix output that can easily be used for visualization of the individual data points in the set. However, the results from these analyses can in fact be used to create trees of their own using hierarchical clustering.

The Unweighted Pair Group Method Arithmetic Mean (UPGMA) algorithm will be used to perform the hierarchical clustering of the analysis methods previously stated [30]. The UPGMA algorithm works to recursively combine data points in set based on their distance matrix. At each step, the pair of objects with the smallest distance is selected and combined. Then, the distances to the rest of the objects and other clusters are recalculated. As shown in equation 2, where the distance between clusters A and B (having the cardinality of $|A|$ and $|B|$ is computed to be the average of all distances $d(x, y)$ between x, y pairs of objects in sets A and B. Basically, the mean distance between the elements.

$$\frac{1}{|A| \cdot |B|} \sum_{x \epsilon A} \sum_{y \epsilon B} d(x, y) \tag{2}$$

The output of the MDS and k-Means clustering algorithms will yield a dataset from which distance matrices can be calculated. These distances matrices will serve as the inputs for the UPGMA algorithm, which will then generate trees based on

that input. It will be of interest to compare the results from the UPGMA-generated trees from the machine learning/statistical methods versus the phylogenetic trees generated from PAUP or other software. We may see areas of agreement in the trees, which will imply support to the validity of the migration model in that area. For areas of disagreement between the trees, more deep-dive analyses can be performed to understand the differences.

Hierarchical clustering (such as UPGMA and others) is available as an R function in the *stats* packages called *hclust* [25].

3.3 Visualization

For all of the machine learning and statistical methods such as multidimensional scaling, clustering, etc., R will be the tool of choice [25]. *Ggplot2* [37] or the R API to Plot.ly [24] will likely be the visualization packages that I will use to generate any visuals surrounding the machine learning and statistical analyses.

In addition, the accepted phylogeny will need to be plotted geographically as were the previous Bantu migration paths from the previous publications. The *Supramap NVector* visualization software created by Janies et al. [17] will be used to plot the dendrograms geospatially, therefore resulting in a truly mapped migration path. The NVector software will render a 3-dimensional globe overlaid with the phyogenetic tree of the proposed Bantu migration. See figure 11, for example. To render 2-dimensional maps (comparable to the maps in figure 1), I will likely use *Leaflet* [1], a Javascript mapping library that has connectors into R [8], or *Tableau* [29]. All three mapping tools provide immersive and interactive functionality for data exploration.

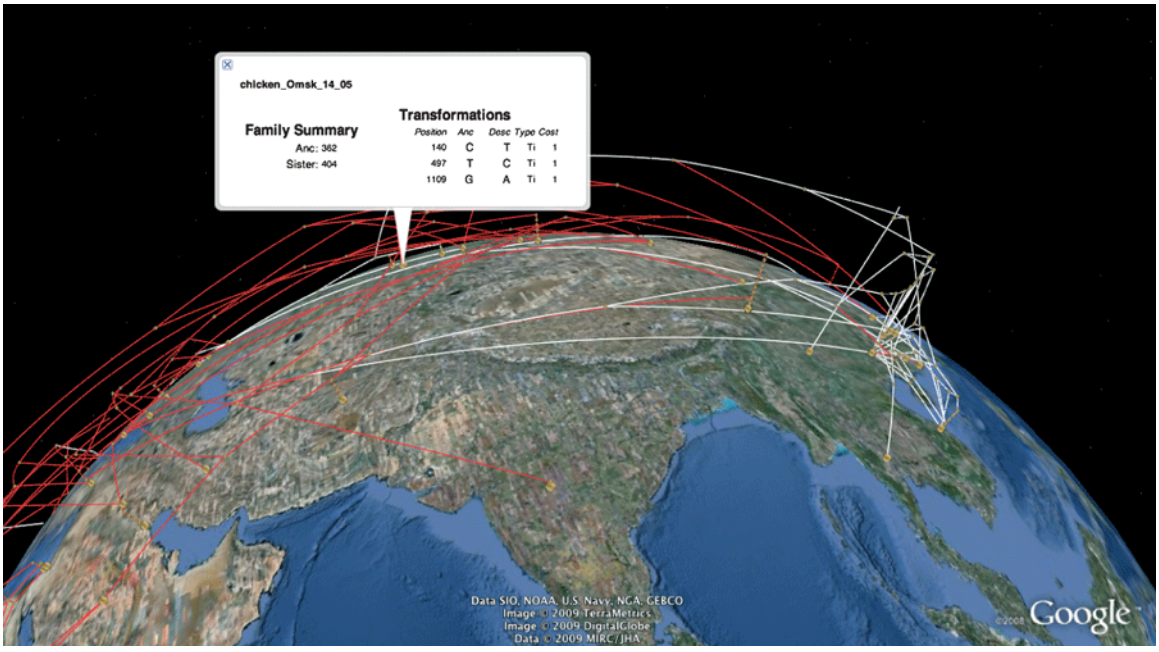

Figure 11: Sample visualization by the *Supramap NVector* software of the spread of the H5N1 avian flu [17].

CHAPTER 4: CONCLUSION AND SIGNIFICANCE

Understanding the evolutionary relationships of a set of organisms is often a difficult problem as getting enough data to prove a migratory path is tricky. For humans, the task is even less straightforward because of the different aspects we pass along to our offspring. We not only pass down out genetic information, but the language(s) we speak as well as the cultural norms we practice. Linguistic and cultural traits can be treated as seriously as genes given their variability and tendency to blend with one another. For each of us, we not only inherited a mixture of genetic traits from our mothers and fathers, but were also taught how to speak and behave by them as well.

Beyond simply creating a combined phylogeny, I plan to test the data in multiple ways. Using machine learning and statistical validation techniques, support for (or rejection of) the migratory model will follow. By looking at the data from multiple mathematical fronts, confidence can be built around the migratory model accuracy. In other words, this will portray confidence that the model follows the true Bantu migration path more closely that previous methods' models.

Visualization will prove to be imperative in the analysis process as well. Being able to plot the various phylogenies, both from the singular and combined datasets, will be necessary in determining the agreement or disagreement to the other, previously-published models. Plus, representing the trees geographically will be the final step in generating my assessment of the migratory path. In addition, the results of the

machine learning methods will be graphically represented for easier interpretation and comparison to the phylogenetic analyses, too.

All steps of this proposal outline my planned approach to test my hypothesis that the Bantu expansion can be characterized by an early primary split in lineages. Depending on the results of the analysis, the hypothesis may be overwhelmingly supported or may suggest an entirely new migration path that differs from any other model to date. The completion of this research will show the versatility and usefulness of integrated data, phylogenetics, machine learning, and visual analytics in bioinformatics and computational biology, specifically in the examination of Bantu migration.

REFERENCES

[1] V. Agafonkin. Leaflet. 2017.

[2] E. Alpaydin. *Introduction to Machine Learning.* MIT Press, 3 edition, 2014.

[3] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. *Technical Report*, 668, Apr 2005.

[4] Y. Bastin, A. Coupez, and M. Mann. *Continuity and Divergence in the Bantu Languages: Perspectives from a Lexicostatistic Study. Annales du Musé Royal de lAfrique Centrale, Sciences Humaines, #162.* Tervuren, 1999.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[6] R. Butler. Congo deforestation, Jan 2016.

[7] J. J. Butt. *The Greenwood dictionary of world history.* Greenwood Press, 2006.

[8] J. Cheng, B. Karambelkar, and Y. Xie. *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*, 2017. R package version 1.1.0.

[9] T. E. Currie, A. Meade, M. Guillon, and R. Mace. Cultural phylogeography of the bantu languages of sub-saharan africa. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1762), 2013.

[10] C. de Filippo, K. Bostoen, M. Stoneking, and B. Pakendorf. Bringing together linguistic and genetic evidence to test the bantu expansion. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1741):3256–3263, 2012.

[11] K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[12] P. A. Goloboff, J. S. Farris, and K. C. Nixon. Tnt, a free program for phylogenetic analysis. *Cladistics*, 24(5):774–786, 2008.

[13] I. Gonzlez and S. Djean. *CCA: Canonical correlation analysis*, 2012. R package version 1.2.

[14] J. P. Gray. Ethnographic atlas codebook. *World Cultures*, 10(1):86136, 1998.

[15] R. Grollemund, S. Branford, K. Bostoen, A. Meade, C. Venditti, and M. Pagel. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, 112(43):13296–13301, 2015.

[16] M. Guthrie. *Comparative Bantu : an introduction to the comparative linguistics and prehistory of the Bantu languages*. Brookfield, VT: Gregg, 1967-1971.

[17] D. A. Janies, T. Treseder, B. Alexandrov, F. Habib, J. J. Chen, R. Ferreira, . atalyrek, A. Varn, and W. C. Wheeler. The supramap project: linking pathogen genomes with geography to fight emergent infectious diseases. *Cladistics*, 27(1):61–66, 2011.

[18] K. Katoh. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):30593066, 2002.

[19] G. Kraemer. *dimRed: A Framework for Dimensionality Reduction*, 2017. R package version 0.0.3.

[20] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, Mar 1964.

[21] S. P. Lloyd. Least square quantization in pcm. *Bell Telephone Technical Memorandum*, 1957.

[22] S. P. Lloyd. Least square quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[23] D. R. Maddison, D. L. Swofford, W. P. Maddison, and D. Cannatella. Nexus: An extensible file format for systematic information. *Systematic Biology*, 46(4):590–621, 1997.

[24] Plotly. Plot.ly. 2017.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[26] A. C. Rencher. *Methods of multivariate analysis*. Wiley, 1 edition, 1995.

[27] T. Russell, F. Silva, and J. Steele. Modelling the spread of farming in the bantu-speaking regions of africa: An archaeology-based phylogeography. *PLOS ONE*, 9(1):1–9, 01 2014.

[28] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[29] T. Software. Tableau. 2017.

[30] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.

[31] A. Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[32] D. L. Swofford. Paup (phylogenetic analysis using parsimony). *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, page 14551455, 2008.

[33] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[34] P. Verdu, E. M. Jewett, T. J. Pemberton, N. A. Rosenberg, and M. Baptista. Parallel trajectories of genetic and linguistic admixture in a genetically admixed creole population. *Current Biology*, 27(16):2529 – 2535.e3, 2017.

[35] W. C. Wheeler, N. Lucaroni, L. Hong, L. M. Crowley, and A. Varón. POY version 5.0. American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php, 2013.

[36] P. Whiteley, W. Wheeler, and M. Xue. Revising the Bantu tree: Words as sequences 2.0. *Current Anthropology*, 2017. in review.

[37] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.