AN INTEGRATED PHYLOGEOGRAPHIC ANALYSIS OF THE BANTU MIGRATION

by

Colby Tyler Ford

A dissertation 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

2018

Approved by:

Dr. Daniel Janies

Dr. Xinghua Shi

Dr. Anthony Fodor

Dr. Mirsad Hadzikadic

Dr. Matthew Parrow

ABSTRACT

COLBY TYLER FORD. An Integrated Phylogeographic Analysis of the Bantu Migration. (Under the direction of DR. DANIEL JANIES)

"Bantu" is a term used to describe lineages of people in around 600 different ethnic groups on the African continent ranging from modern-day Cameroon to South Africa. The migration of the Bantu people, which occurred around 3,000 years ago, was influential in spreading culture, language, and genetic traits and helped to shape human diversity on the continent. Research in the 1970s was completed to geographically divide the Bantu languages into 16 zones now known as "Guthrie zones" [25].

Researchers have postulated the migratory pattern of the Bantu people by examining cultural information, linguistic traits, or small genetic datasets. These studies offer differing results due to variations in the data type used. Here, an assessment of the Bantu migration is made using a large dataset of combined cultural data and genetic (Y-chromosomal and mitochondrial) data.

One working 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 what is now called the Congolese forest (i.e. "early split"). A competing hypothesis is that the split occurred south of the forest (i.e. "late split").

Using the comprehensive dataset, a phylogenetic tree was developed on which to reconstruct the relationships of the Bantu lineages. With an understanding of these lineages in hand, the changes between Guthrie zones were traced geospatially.

Evidence supporting the "early split" hypothesis was found, however, evidence for

several complex and convoluted paths across the continent were also shown. These findings were then analyzed using dimensionality reduction and machine learning techniques to further understand the confidence of the model.

ACKNOWLEDGMENTS

First off, I would like to thank my advisor, Dr. Daniel Janies, for his amazing wisdom during the course of this research. Sincere gratitude to my committee members, Dr. Xinghua Shi, Dr. Anthony Fodor, Dr. Mirsad Hadzikadic, and Dr. Matthew Parrow for their valuable support and guidance.

Special thanks to the Wayland H. Cato Jr. Doctoral Fellowship, which provided generous financial support during my time as a doctoral student.

Finally, and most importantly, I would like to thank both of my grandmothers. From the time I was small, these two strong women set the foundation for my academic success by always supporting my curiosity, caring about my intellectual achievement, and placing a strong importance on learning. And for these things, I thank you with you all my heart. Without you two, I would not have made it this far. I love you both.

TABLE OF CONTENTS

LIST OF FIGURES ix

LIST OF TABLES xii

CHAPTER 1: INTRODUCTION AND BACKGROUND 1

1.1. Current Migratory Models 2

1.2. The Early Split Hypothesis 5

1.2.1. The Role of Geographic Barriers in the Bantu Migration 5

CHAPTER 2: PHYLOGENETIC ANALYSIS USING HETEROGENEOUS DATA 8

2.1. Introduction 8

2.2. Materials and Methods 9

2.2.1. Data Curation 9

2.2.2. Data Workflow 14

2.2.3. Tree Search 15

2.2.4. Reconstructing Migration Trajectory 19

2.3. Results 21

2.3.1. Trees 21

2.3.2. Migratory Path Visualization 24

2.4. Discussion 31

CHAPTER 3: COMPARING TAXA PROXIMITIES USING DIMENSIONALITY REDUCTION 36

3.1. Introduction 36

3.2. Materials and Methods 37

3.2.1. Data Shaping 37

3.2.2. Multidimensional Scaling 41

3.2.3. Laplacian Eigenmaps 43

3.3. Results 45

3.4. Discussion 63

CHAPTER 4: ASSESSMENT AND COMPARISON OF MODEL ACCURACY USING MACHINE LEARNING 65

4.1. Introduction 65

4.2. Materials and Methods 67

4.2.1. Random Forest Model Generation 68

4.3. Results 70

4.4. Discussion 80

CHAPTER 5: CONCLUSION 81

5.1. New Working Hypothesis 83

5.1.1. Evidence of Early Split 83

5.1.2. Evidence of Subsequent Migration 84

5.2. Significance and Future Work 84

5.2.1. Applicability to Other Research 86

REFERENCES 88

APPENDIX A: DATABASE SETUP CODE 93

APPENDIX B: YCHR AND CULTURAL DATASET CHARACTER POSITION LISTS 97

APPENDIX C: DIMENSIONALITY REDUCTION CODE 103

APPENDIX D: MACHINE LEARNING DATA PREPARATION CODE 106

APPENDIX E: LINKS TO RESEARCH MATERIALS 108

LIST OF FIGURES

FIGURE 1: Hypotheses of Bantu language expansion. a) "early split" vs. b) "late split" de Filippo et al., 2011 [16], (fig. 2) c) Currie et al., 2013 [14], (fig. 2b) d) Grollemund et al., 2015 [24], (fig. 2A) main nodes and branches. Note that these paths are summarized to the Guthrie zone, which may be at a higher level than the published work's results. 4

FIGURE 2: Spatial distribution of the African rainforests derived from MODIS data. (Butler, 2016 [10]) 7

FIGURE 3: Mapped locations of the 138 genetic samples. Note that multiple samples may have come from the same location. 10

FIGURE 4: Tree generation and migratory model creation workflow. 15

FIGURE 5: Parsimonious phylogenetic tree generated by POY using the combined data and the haversine distance-based Sankoff matrix. 22

FIGURE 6: Parsimonious phylogenetic tree generated by POY using the mtDNA data and the haversine distance-based Sankoff matrix. 23

FIGURE 7: Parsimonious phylogenetic tree generated by POY using the Ychr data and the haversine distance-based Sankoff matrix. 25

FIGURE 8: Parsimonious phylogenetic tree generated by POY using the cultural data and the haversine distance-based Sankoff matrix. 26

FIGURE 9: Migratory model generated using the Guthrie zone transitions from the combined data and the haversine distance-based Sankoff matrix. 27

FIGURE 10: Migratory model generated using the Guthrie zone transitions from the mtDNA data and the haversine distance-based Sankoff matrix. 29

FIGURE 11: Migratory model generated using the Guthrie zone transitions from the Ychr data and the haversine distance-based Sankoff matrix. 30

FIGURE 12: Migratory model generated using the Guthrie zone transitions from the cultural data and the haversine distance-based Sankoff matrix. 32

FIGURE 13: All four generated migratory models overlaid for comparison. 34

FIGURE 14: Dimensionality Reduction data shaping workflow. 40

FIGURE 15: 2D non-metric MDS scatter plot of inner joined data summarized by TaxaID. 46

FIGURE 16: A zoomed-in view of figure 15, disregarding the farthest two points (S33Sotho and M52Lala). 47

FIGURE 17: 2D non-metric MDS scatter plot of outer joined data summarized by TaxaID. 49

FIGURE 18: A zoomed-in view of figure 17 of the densest cluster of points. 50

FIGURE 19: 2D non-metric MDS scatter plot of inner joined data summarized by Guthrie zone. 51

FIGURE 20: A filtered view of figure 19, disregarding Guthrie zone M. 52

FIGURE 21: 2D non-metric MDS scatter plot of outer joined data summarized by Guthrie zone. 53

FIGURE 22: A filtered view of figure 21, disregarding Guthrie zones A, F, and M. 54

FIGURE 23: 2D LE scatter plot of inner joined data summarized by TaxaID. 55

FIGURE 24: A zoomed-in view of figure 23, disregarding the farthest two points (K11Ciokwe and R111Umbundu). 56

FIGURE 25: 2D LE scatter plot of outer joined data summarized by TaxaID. 57

FIGURE 26: A zoomed-in view of figure 25, disregarding the farthest two points (C55Kele and M54Lamba). 58

FIGURE 27: 2D LE scatter plot of inner joined data summarized by Guthrie zone. 59

FIGURE 28: 2D LE scatter plot of outer joined data summarized by Guthrie zone. 60

FIGURE 29: Map of TaxaID locations, colored by Guthrie zone. 61

FIGURE 30: Example Azure Machine Learning experiment workflow. 68

FIGURE 31: Confusion matrix of Guthrie zone predictions versus actual zones by model. 72

FIGURE 32: Box plots of accuracy by model, colored by Guthrie zone. 74

FIGURE 33: Box plots of accuracy by model, colored by Guthrie zone. Shown excluding zones E, F, G, and N, which had low observation counts in the training datasets for these models. 75

FIGURE 34: Box plots of additional model metrics, colored by Guthrie zone. 76

FIGURE 35: Box plots of accuracy by model, colored by cross-validation fold. 78

FIGURE 36: Box plots of additional model metrics, colored by cross-validation fold. 79

FIGURE 37: Color-coded migration directions between Guthrie zones. 85

LIST OF TABLES

TABLE 1: Number of individual samples and the representation of Guthrie zones for each dataset. 11

TABLE 2: Bibliography of data sources by type. 13

TABLE 3: Calculated centroids of each Guthrie zone. 17

TABLE 4: Haversine distance-based cost matrix used in the parsimonious tree search. 18

TABLE 5: Sample structure of the migratory model dataset. Note this is the data that allows for the visualization of the migratory paths for all the models, both from this work and from previous models. 20

TABLE 6: Schema of the mtDNA database table. 38

TABLE 7: Schema of the Ychr database table. 38

TABLE 8: Schema of the Cultural database table. 38

TABLE 9: Schema of the Geographic Information database table. 39

TABLE 10: Stress values of the MDS Analyses. Note that all are below 0.05. 62

TABLE 11: Optimal k parameters in k-NN step for LE Analyses and their corresponding k-NN model accuracies. 62

TABLE 12: Input parameters for the Azure Machine Learning Multiclass Decision Forest module. 69

TABLE 13: Input parameters for the Azure Machine Learning Partition and Sample module 70

TABLE 14: Overall random forest 3-fold cross-validation accuracy by model. 71

TABLE 15: Zone coverage in the Random Forest model training datasets. 73

TABLE 16: R output of the 5-proportion Z-test for equivalence of model accuracies. 77

TABLE 17: Y-Chromosome STR Loci List, adapted from Butler et al., 2008 [9] 97

TABLE 18: Cultural Ethnographic Atlas Question List, from Gray et al., 1998 [23] 98

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 200 million people from Central and Southern Africa. Among the Bantu population, there are around 600 languages (including dialects) spoken. It is thought that the Bantu people originated from what is now Cameroon ˜3,000 years ago and then spread to the east and south of the continent [11]. However, the exact migratory or expansion path that was taken is unknown and, as a result, the point of some debate.

The Bantu migration was a majorly influential spreading of culture, language, and of course, genetic traits. The Bantu people had a more sedentary and settled lifestyle than that of the indigenous people. Specifically, the Bantu placed a larger importance on farming, whereas the indigenous people were commonly forest foragers. With developments such as agricultural technology, the making of ceramics, and the use of iron, new advancements drove the expansion to new ecological zones for the use of the land's resources [42, 19]. Thus, the spread of these individuals significantly shaped the human diversity on the continent.

The Bantu expansion has been an active area of research for over 50 years. The work of researchers Malcom Guthrie, in the 1960s and 1970s [26], and Roland Oliver, in the late 1960s [43], helped shape the initial understanding of the Bantu people, their expansion, and its importance. Guthrie's work resulted in a linguistic division

of the region into 16 Guthrie zones. Oliver laid the foundation of this area of research when he created one of the first expansion models, which consisted of four distinct phases and was based on archaeological and linguistic data coupled with geographical features of the land.

By looking at changes in the Bantu people geographically, the trajectory by which the individuals moved around the continent throughout time can be approximated. 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 overall agreement that the group started in what is now the Republic of 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 anthropologically-driven approach. One of these approaches includes using linguistic data, 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 [24]. Other research has surfaced that suggests that

the hypothesis concerning Bantu expansion can be tested by using both linguistic and some genetic data [16]. 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 [46] across the continent.

Some of these hypotheses are built on the notion that the likelihood by which a group will have to adapt in the new environment affects the migratory choices that the group will make. For example, there is a notion that humans preferred to migrate through habitats that were most similar in climate, terrain, vegetation, or habitat compared to where they had previously lived (referred to as a "path of least resistance" approach). This resulted in a later theory known as the "beachcomber" hypothesis, which proposes that the individuals may have migrated using a more coastal route [40]. However, given the data from the previous migratory model publications, this does not seem to be the case. Large movements into vastly different areas of the continent are shown in each of the different models which do not support this theory.

Having this belief of a "path of least resistance" may have increased the risk of skewed interpretations in the previous migratory research. Incorrectly interpreting messy results risks apophenia. That is, perceiving results in a certain way based on the prior hypothesis, rather than from an objective viewpoint. With this in mind, the analyses in this research were performed with little dependence on the aforementioned notions.

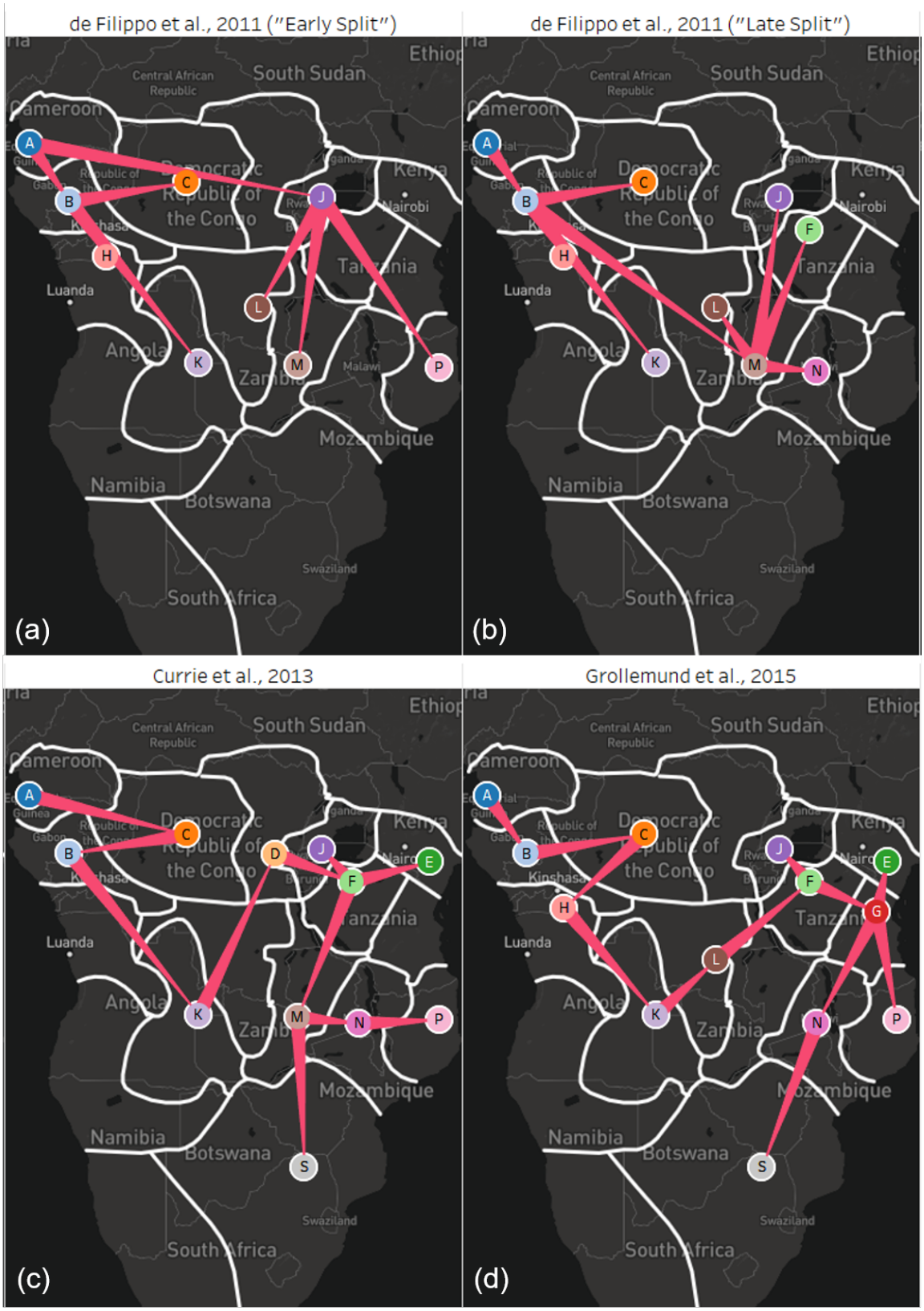

Figure 1: Hypotheses of Bantu language expansion.
a) "early split" vs. b) "late split" de Filippo et al., 2011 [16], (fig. 2)
c) Currie et al., 2013 [14], (fig. 2b)
d) Grollemund et al., 2015 [24], (fig. 2A) main nodes and branches. Note that these paths are summarized to the Guthrie zone, which may be at a higher level than the published work's results.

1.2 The Early Split Hypothesis

The working hypothesis behind this research 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, the divergence and migration should flow from north to south as well as east to west. 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 Cameroon, correlated to their geographic distance. This is opposite of the "path of least resistance" theory where the thought is that east-west migrations should have occurred more than north-south migrations because the former is less likely to result in encountering variations in climate or habitat [18]. Also, conversely to the de Filippo models, it is expected that the migratory path will take a less linear route, but spread in a branching manner.

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

1.2.1 The Role of Geographic Barriers in the Bantu Migration

In any migratory scenario, there could be limitations as to the path a group of organisms can take. For example, without proper technology such as large ships

or the ability to fly, it is unlikely that the Bantu people migrated via sea or air. Therefore, the only way this group expanded was on land.

Even on land, though, there are regions of the continent that would have been more difficult to traverse than others. This likely shaped the migration path by limiting the possibly trajectories people could have traveled.

One such region is the Congolese forest. This area of the continent is considerably more wet and verdant than the surrounding, more arid areas. Certain parts of this rainforest are conducive to farming and habitation, whereas other parts are more swamp-like and therefore less inhabitable or more difficult to traverse.

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

Additionally, there are large mountain ranges on the eastern side of the continent, such as the Eastern Rift mountains, that may have also played a role in shaping the Bantu migration trajectory. Mountains in these ranges, such as Kilimanjaro, are over 5,000 meters high and have very difficult climates on top [**?**]. These regions could have been treacherous to surmount, therefore limiting the spread the Bantu people in these areas.

If the early split hypothesis is true and the Bantu expansion went around the areas such as the Congolese forest or the Eastern Rift, further research is needed to understand if this was truly due to the geographical features such as mountains and rainforests prohibiting migrant travel and settling or if other factors played a role

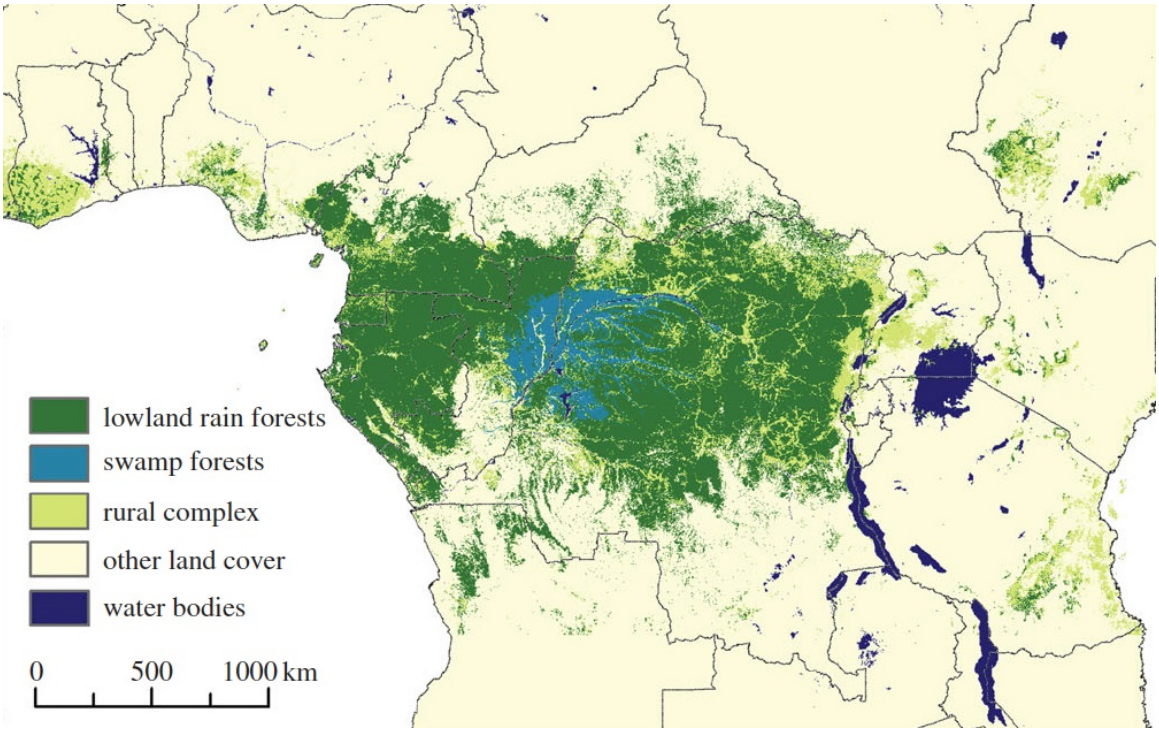

Figure 2: Spatial distribution of the African rainforests derived from MODIS data. (Butler, 2016 [10])

instead.

CHAPTER 2: PHYLOGENETIC ANALYSIS USING HETEROGENEOUS DATA

2.1 Introduction

To begin analyzing the migratory path of the Bantu population, a model tree is generated using a parsimonious phylogenetic analysis. Typically, phylogenetic analysis is performed on a single data type, such as a set of DNA sequences or amino acid sequences. However, the goal was to use heterogeneous data together to create a more complete picture of the Bantu migration. The use of Y-chromosomal, mitochondrial DNA, and cultural data together was presumed to create a phylogenetic tree which is better at explaining the multifaceted nature of the variation among the Bantu groups. These three datasets look to represent the paternal, maternal, and cultural lineages of the people.

Employing traditional methods of parsimonious tree generation does not work without careful manipulation of the input data, along with the combination file generation that contains all the data together, rather than three individual inputs.

Given the variety of the data, certain data processing exercises must be completed before the data are acceptable for use in tree generation. The required processing steps differ depending on the input data that are available.

2.2 Materials and Methods

2.2.1 Data Curation

Genetic data representing the paternal lineage of the Bantu groups has been curated using Y-chromosomal (Ychr) short tandem repeats (STRs). The STR markers are published in a .nexus file as integers which correspond to the number of tandem repeats the individual possesses for a given microsatellite. The data are then re-coded to retain the representation of the allelic makeup as a character string. This representation allows a compact display of the STRs as a 12-character string where each character corresponds to one STR marker, and the STR name is given as a character state label. The order of the character state labels is as follows: 1: DYS389_I, 2: DYS389_II, 3: DYS385a, 4: DYS385b, 5: DYS391, 6: DYS390, 7: DYS393, 8: DYS392, 9: DYS19, 10: DYS437, 11: DYS438, and 12: DYS439. See table 17 for more information [9]. This data contains 1,724 Bantu-speaking STR profiles and 157 non-Bantu Nigerian samples (to serve as an outgroup set) for a total of 1,881 samples. This data covers 49 distinct language groups (plus 2 outgroups), which represents 11 of the 16 Guthrie zones. This data is provided in the .nexus format.

In addition to representing the paternal lineage, mitochondrial DNA samples are also collected to represent the maternal lineage. This data includes samples from 742 Bantu-speaking individuals and an additional 4 non-Bantu Nigerian samples (to serve as an outgroup set) for a total of 746 samples. The mtDNA data covers 56 distinct language groups (plus 2 outgroups), which represents 12 of the 16 Guthrie zones. This data is provided in an unaligned .fasta format.

For some previous Bantu migration publications, 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 [16], 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 the data in this work, the sampling locations from each of the papers were mapped. See figure 3. The spread of the Ychr and mtDNA genetic samples appears to be more evenly distributed compared to other publications' datasets, at least from a purely visual inspection.

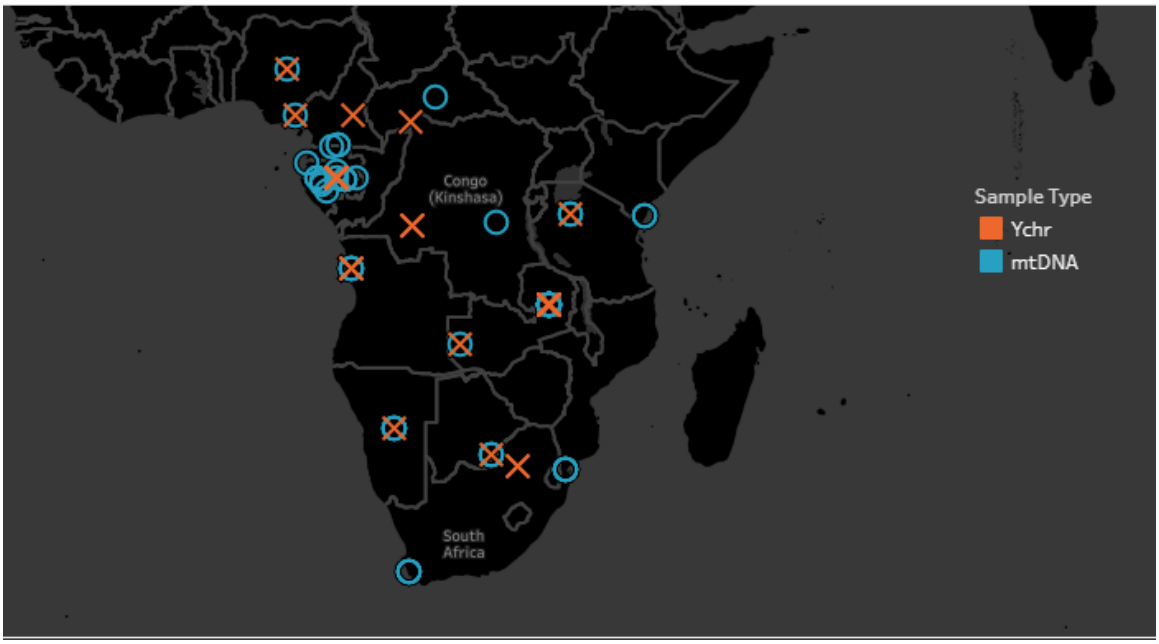

Figure 3: Mapped locations of the 138 genetic samples. Note that multiple samples may have come from the same location.

In addition to the genetic data, cultural data has been collected and is used in conjunction with the mtDNA and Ychr data. The cultural data is composed of 92 cultural traits outlined by the Ethnographic Atlas Codebook [23]. These "traits" are cultural customs, such as the dependence on hunting, gathering, and fishing, as well

Table 1: Number of individual samples and the representation of Guthrie zones for each dataset.

Guthrie zone	**mtDNA**	**Ychr**	**Cultural**	**Total**
A	5	155	5	**165**
B	10	569	5	**584**
C	4	215	11	**230**
D			5	**5**
E	2		6	**8**
F		30	3	**33**
G	2		7	**9**
H	21	49	6	**76**
J			9	**9**
K	184	165	3	**352**
L	42	36	2	**80**
M	104	118	6	**228**
N	8	7	6	**21**
P			2	**2**
R	148	173	3	**324**
S	212	207	12	**431**
Z (Outgroup)	4	157	12	**173**
Total	**746**	**1881**	**103**	**2730**
Zone Coverage	*75%*	*69%*	*100%*	*100%*

as marriage practices, the use of animals for food and agriculture, and gender roles. For a full list of cultural traits collected from the Ethnographic Atlas Codebook, see table 18. In the cultural data, 93 language groups are represented, which includes all 16 Guthrie zones.

Only 16 of the groups (TaxaIDs) are exact matches in between all three of the datasets. However, despite having a low number of specific groups matching, a majority of the Guthrie zones are represented in each dataset. Zones D, E, F, G, J, and P are underrepresented in the genetic data only. All zones are represented in the cultural data. See table 1. The migratory model is generated at the Guthrie zone level. So, this coverage is sufficient for that analysis.

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

Table 2: Bibliography of data sources by type.

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.[1]
Mitochondrial	Behar, Doron M., et al. "The dawn of human matrilineal diversity." The American Journal of Human Genetics 82.5 (2008): 1130-1140.[2]
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.[12]
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. [17]
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).[22]
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.[7]
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).[41]
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.[44]
Y-chromosomal	Fujihara et al. "Allele frequencies and haplotypes for 28 Y-STRs in Ovambo population." Legal Medicine. 11 (2009): 205-208.[21]
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.[4]
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.[13]
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.[17]
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.[27]
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.[37]
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.[38]
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.[51]
Cultural	Gray, J. P., "Ethnographic Atlas Codebook" World Cultures 10.1 (1998): 86-136.[23]

2.2.2 Data Workflow

Given the diversity of the data, each dataset must be standardized such that each can be similarly analyzed, both separately and combined. To integrate the datasets for a combined analysis, each dataset must be converted to the same format so that the integration can occur. The optimal format for both standalone and combined analyses is the .nexus file format. Nexus is an extensible file format that is popular in the bioinformatics field as it is flexible enough to house different types of data and metadata [39].

The Ychr 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 (Multiple Alignment using Fast Fourier Transform) [32] (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 obtained pre-formatted as a .nexus file. No changes were made to the cultural data's format.

Once the data conversions are complete, the individual blocks of sequences in the separate .nexus files are then copied into a single, combined file. This workflow now results in 4 .nexus files, one for each of the separate datasets and one for the combined data. See figure 4 for a graphical representation of the workflow.

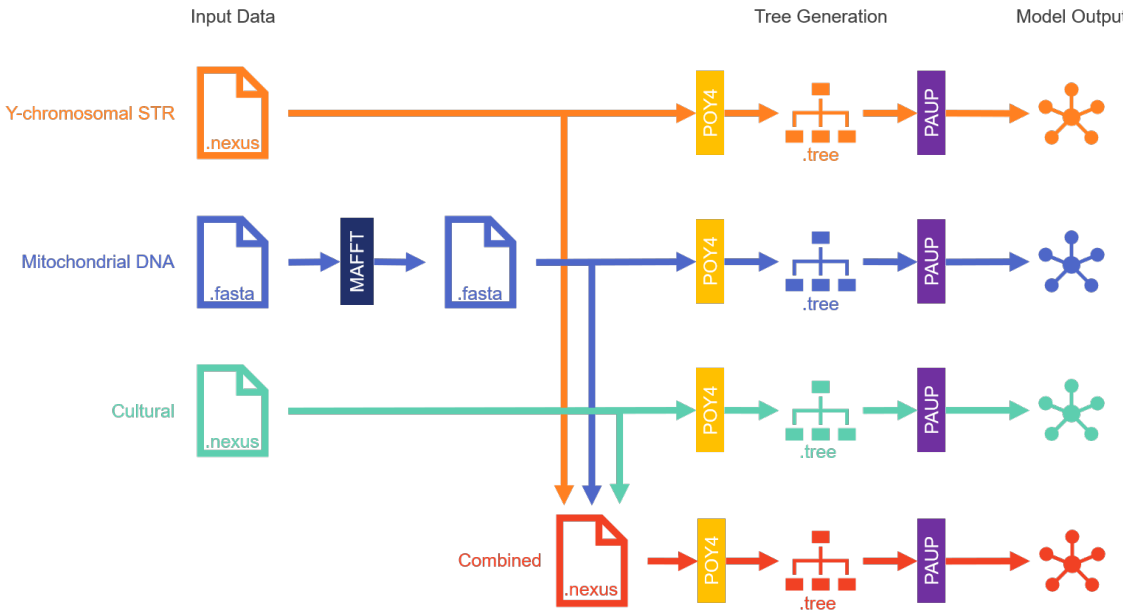

Figure 4: Tree generation and migratory model creation workflow.

2.2.3 Tree Search

Parsimony methods to building phylogenetics trees are some of the most widely used algorithms in the field. These methods work by finding the tree that can explain the sequences in the observed data using the least number of substitutions [20].

To identify the "best" tree, the parsimony algorithm searches through all topologies and assigns a cost to each. In other words, there are two distance steps for parsimonious tree generation:

1. Compute the cost of a given tree T.

2. Search through all trees and find the overall minimum cost.

With the .nexus files as inputs, POY4 [54] was used to perform the tree search. For each of the analyses (three separate and one combined), a single, heuristically-derived

consensus tree is created using the parsimonious method. The resulting tree is the one found to have the lowest cost, or number of substitutions given the input data.

2.2.3.1 Haversine Distances as a Cost Function

Parsimonious tree reconstructions are dependent on a set of cost parameters known as a cost matrix or a Sankoff matrix [47]. A matrix can be derived using various methods, such as the probability from transitioning from state i to j or some logic around adjacent versus non-adjacent geographic transitions.

Here, the haversine distance was selected as the metric from which to generate the cost matrix for the tree reconstruction. Haversine distance is the distance between two points on a sphere and is used as a method to calculate the distance between two locations on the planet [8]. In this scenario, the geolocation of individual Guthrie zones is important in the overall postulation of the migration path. Given the availability of Guthrie zone and taxa locations, this distance is easily calculated.

To calculate the haversine distance d on a sphere of radius R, the steps are as follows [33]:

$$a = \sin^2(\frac{lat_1 - lat_2}{2}) + \cos(lat_1) \cdot \cos(lat_2) \cdot \sin^2(\frac{long_1 - long_2}{2})$$

$$c = 2 \arctan 2(\sqrt{a}, \sqrt{1-a})$$

$$d = R \cdot c$$

For these calculations, the radius of the Earth is calculated as $6371e3$.

Using the latitude and longitude in table 3, the haversine distance was calculated and resulted in the matrix in table 4. Note that the indel cost is the average of the

Table 3: Calculated centroids of each Guthrie zone.

Guthrie zone	**Latitude**	**Longitude**
A	2.348888889	10.31111111
B	-1.753125	13.153125
C	-0.40769231	21.38846154
D	-1.78	27.604
E	-2.35	38.45
F	-3.73333333	33
G	-5.85	37.7125
H	-5.57142857	15.75
J	-1.44444444	30.91111111
K	-12.963	22.225
L	-9.16	26.44
M	-13.1222222	29.19444444
N	-13.5571429	33.51428571
P	-13.3	39.125
R	-15.825	15.45
S	-23.125	29.665625
Z	5.774285714	9.385

non-zero values in the rest of the matrix. This makes the cost to make an insertion or deletion less than making a large jump between distant Guthrie zones, but not so low that non-adjacent zone transitions are never favored over indels.

Table 4: Haversine distance-based cost matrix used in the parsimonious tree search.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**J**	**K**	**L**	**M**	**N**	**P**	**R**	**S**	**Z**	**Indel**
A	0	555	1269	1977	3171	2611	3176	1068	2329	2151	2198	2702	3111	3624	2098	3524	395	1755
B	555	0	928	1606	2812	2215	2762	513	1974	1597	1685	2171	2596	3133	1585	2970	936	1755
C	1269	928	0	708	1909	1342	1910	850	1065	1399	1122	1655	1980	2423	1834	2680	1499	1755
D	1977	1606	708	0	1207	637	1209	1381	370	1377	831	1273	1462	1802	2053	2384	2191	1755
E	3171	2812	1909	1207	0	624	398	2543	844	2140	1529	1572	1359	1220	2931	2497	3352	1755
F	2611	2215	1342	637	624	0	573	1923	344	1567	943	1125	1094	1259	2343	2186	2827	1755
G	3176	2762	1910	1209	398	573	0	2430	900	1873	1296	1235	973	843	2669	2106	3400	1755
H	1068	513	849	1381	2543	1923	2430	0	1744	1086	1244	1696	2139	2702	1141	2457	1446	1755
J	2329	1974	1065	370	844	344	900	1744	0	1599	990	1312	1377	1599	2330	2415	2521	1755
K	2151	1597	1399	1377	2140	1567	1873	1086	1599	0	645	755	1224	1830	796	1376	2521	1755
L	2198	1685	1122	831	1529	923	1296	1244	990	625	0	533	913	1458	1404	1590	2516	1755
M	2702	2171	1655	1273	1572	1125	1235	1696	1312	755	533	0	470	1075	1510	1113	3034	1755
N	3111	2596	1980	1462	1359	1094	973	2139	1377	1224	913	470	0	608	1959	1139	3423	1755
P	3624	3133	2423	1802	1220	1259	843	2702	1599	1830	1458	1075	608	0	2562	1479	3909	1755
R	2098	1585	1834	2053	2931	2343	2669	1141	2330	796	1404	1510	1959	2562	0	1695	2493	1755
S	3524	2970	2680	2384	2497	2186	2106	2457	2415	1376	1590	1113	1139	1479	1695	0	3896	1755
Z	395	936	1499	2191	3352	2827	3400	1446	2521	2521	2516	3034	3423	3909	2493	3896	0	1755
Indel	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755	1755	0

2.2.4 Reconstructing Migration Trajectory

Once the best tree is generated for each of the datasets, the next step is to extract the apomorphies from each tree. Apomorphy is a derived trait, specialized to a particular group [58]. In this case, the derived trait is the transition between one Guthrie zone to another.

By understanding the apomorphic transitions between Guthrie zones for each of the taxa in a tree, a migration trajectory is reconstructed. However, tracing the apomorphies by simply looking a large tree can be difficult if the tree is complex. As such, using a computational method to perform this search speeds up the extraction of this information.

PAUP: Phylogenetic Analysis Using Parsimony is a widely available tool for creating and analyzing phylogenetics trees. This tool was used to extract the apomorphies from the tree file in each analysis. Since the trees were generated using a more robust and flexible tool like POY, PAUP was only used as a secondary extractor of the necessary apomorphy information. This output is a list of the changes in the format "A ==> B, B ==> H, F −− > G" and so on. The double arrow "==>" represents unambiguous changes whereas the single arrow "−− >" represents ambiguous changes. This information is then transferred into a more specialized dataset that contains the relevant attributes for visualization purposed such as latitude, longitude, Guthrie zone, and model information. See table 5.

Table 5: Sample structure of the migratory model dataset. Note this is the data that allows for the visualization of the migratory paths for all the models, both from this work and from previous models.

Path ID	**Point Order**	**Apomorphy Type**	**Latitude**	**Longitude**	**Guthrie Zone**	**Model**	**ModelType**
A-B	1	Unambiguous	2.3489	10.3111	A	Current Work - Combined	Current Work
A-B	2	Unambiguous	-1.7531	13.1531	B	Current Work - Combined	Current Work
B-H	1	Unambiguous	-1.7531	13.1531	B	Current Work - Combined	Current Work
...	...	...	...	...	...	...	...
E-F	2	Ambiguous	-3.7333	33	F	Current Work - Mitochondrial DNA	Current Work
D-L	1	Ambiguous	-1.78	27.604	D	Current Work - Mitochondrial DNA	Current Work
...	...	...	...	...	...	...	...
M-L	1		-13.1222	29.1944	M	de Filippo et al., 2011 ("Late Split")	Reference Work
M-L	2		-9.16	26.44	L	de Filippo et al., 2011 ("Late Split")	Reference Work

2.3 Results

2.3.1 Trees

Four parsimonious phylogenetic trees were generated in POY: a combined tree using all three datasets (figure 5), a tree using only the mtDNA data (figure 6), a tree using only the Ychr data (figure 7) and a tree using only the cultural data (figure 8).

Regarding the tree generated using the combined data, figure 5, lineages from Guthrie zones S, P, F, and R are monophyletic whereas others are not. The pathways in this model are largely unambiguous.

The trees generated using a singular dataset each have the same overall trajectory as the combined tree, but with with slight variations. One such variation includes the presence of more ambiguity in the reconstruction of the ancestral Guthrie zone states.

As for the tree generated using the mtDNA data, figure 6, there is variation from the combined model in that there are movements from zones B to C and a split from E to F and E to G. All of these are ambiguously reconstructed in the mtDNA model, though. In this tree, the only zones in which the Guthrie character maps to a monophyletic group are H, F, E, G, R, P and S.

In terms of variation in results derived from the tree generated using only Y-chromosomal data with that of combined data tree, the main split occurs in zone J leading to M. See figure 7. This split occurs at zone D in the combined model. However, this is ambiguously reconstructed in the Y-chromosomal model. In this tree, the only zones in which the Guthrie character maps to a monophyletic group

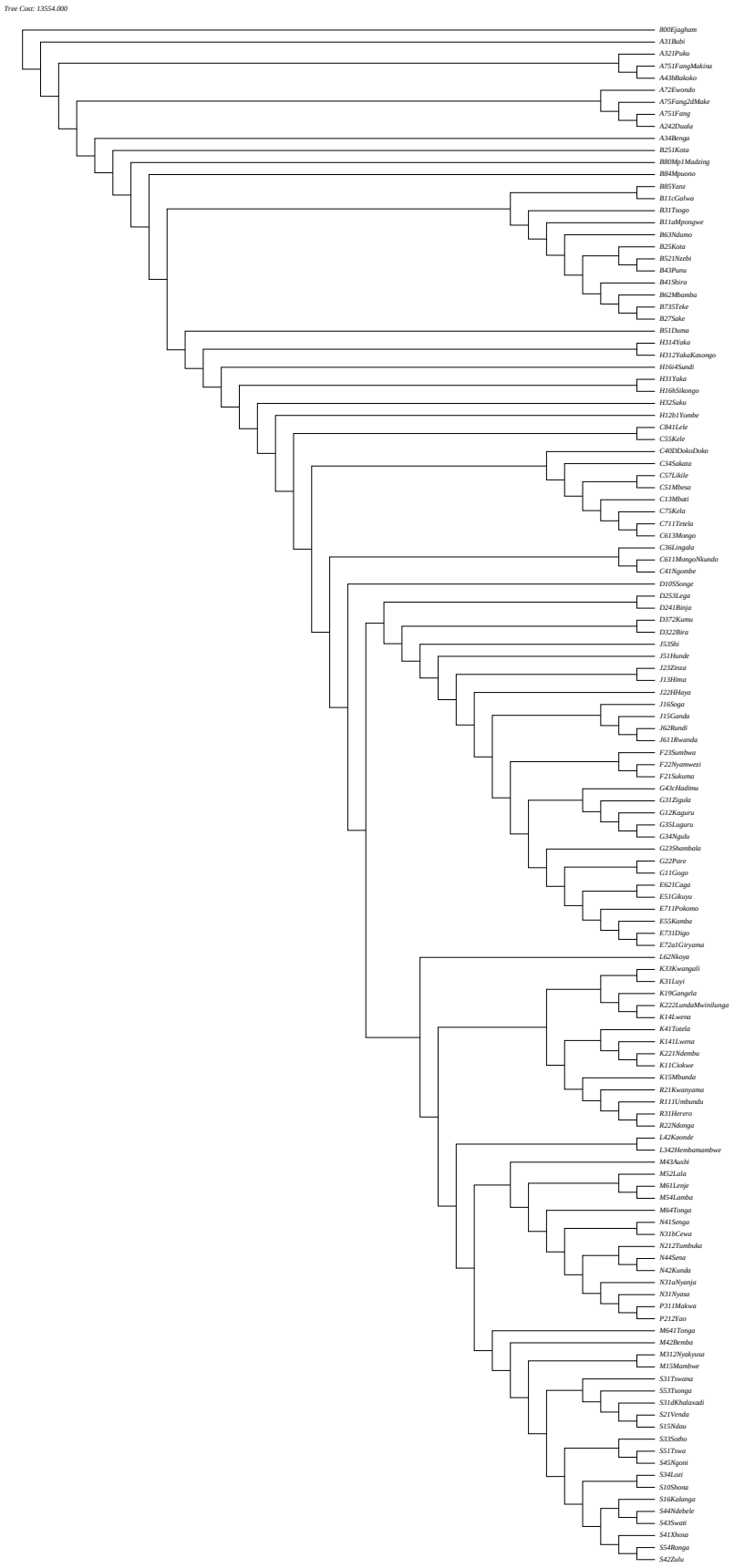

Figure 5: Parsimonious phylogenetic tree generated by POY using the combined data and the haversine distance-based Sankoff matrix.

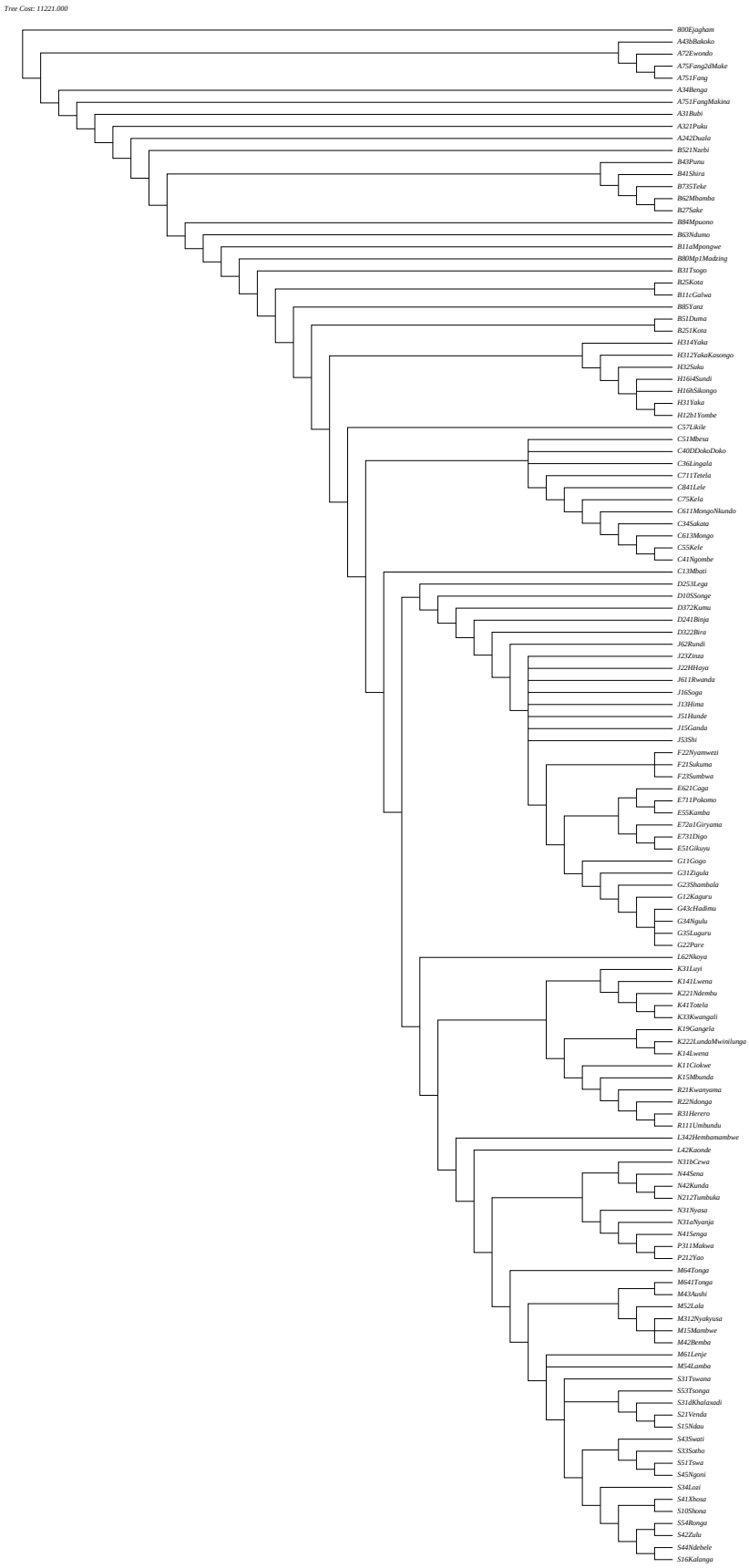

Figure 6: Parsimonious phylogenetic tree generated by POY using the mtDNA data and the haversine distance-based Sankoff matrix.

are C, D, F, E, R, P, and S.

One variation that is seen in the tree generated using only cultural data is the presence of a D to K zone transition, albeit ambiguously reconstructed. See figure 8. In this tree, the only zones in which the Guthrie zone character maps to a monophyletic group are F, L, P, R, and S.

2.3.2 Migratory Path Visualization

Using the .tree files from the POY output, apomorphies were extracted using PAUP. The apomorphic characters, or derived characters/traits that arose directly from the ancestor and is unique to all descendants [58], are treated as transitions between Guthrie zones. These zone transitions are transformed and combined into the data seen in figure 5. From here, Tableau was used to plot these transitions geographically [48]. This tool provides immersive and interactive functionality for data exploration, which was useful in the comparison of the generated models as well as the comparison to the reference models from previous researchers.

For the combined migratory model, the trajectory is seen starting in the northwest in Guthrie zone A and moving southward to zones B and H. Then, the model shows an eastward movement through zones C and D. At zone D a split occurs with migration going east through zones J, F, G, and E and south to zone L. From zone L, another split occurs, heading west to zones K and R and heading southwest to zone M. At zone M, another split occurs, with one branch heading south and ending at zone S and heading east to zone N and ending at P. See figure 9. Note that the only ambiguous transitions occur from zone J to F and F to G. Every other transition is unambiguous.

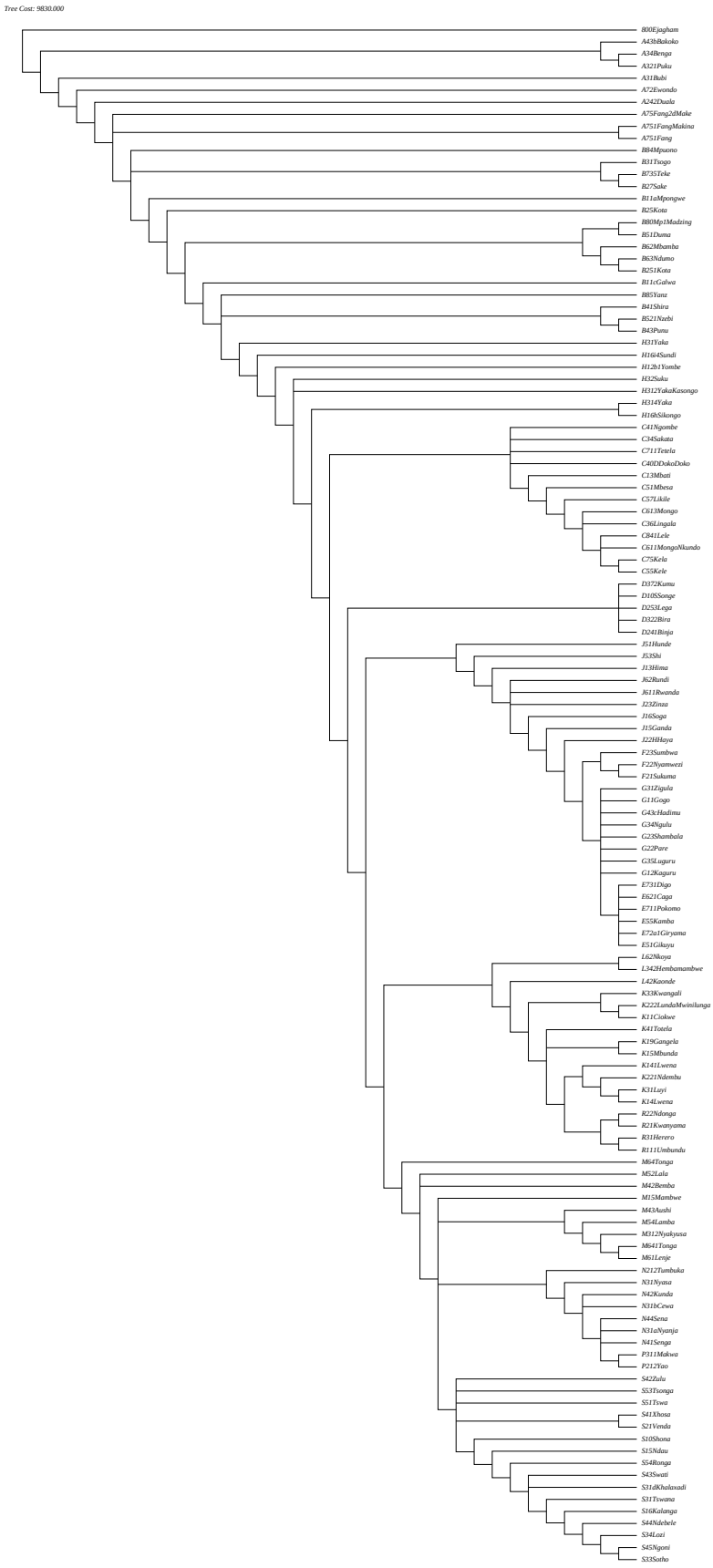

Figure 7: Parsimonious phylogenetic tree generated by POY using the Ychr data and the haversine distance-based Sankoff matrix.

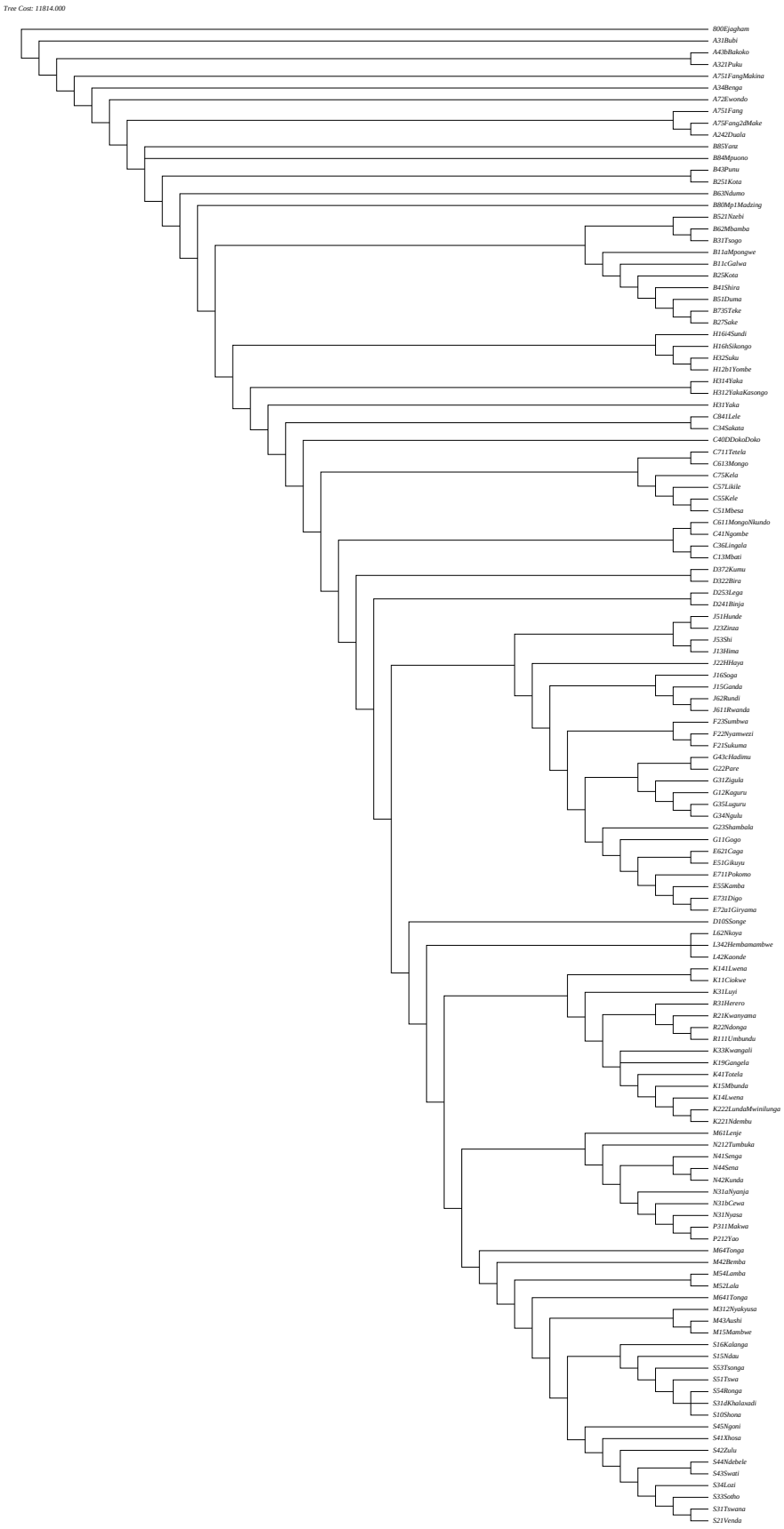

Figure 8: Parsimonious phylogenetic tree generated by POY using the cultural data and the haversine distance-based Sankoff matrix.

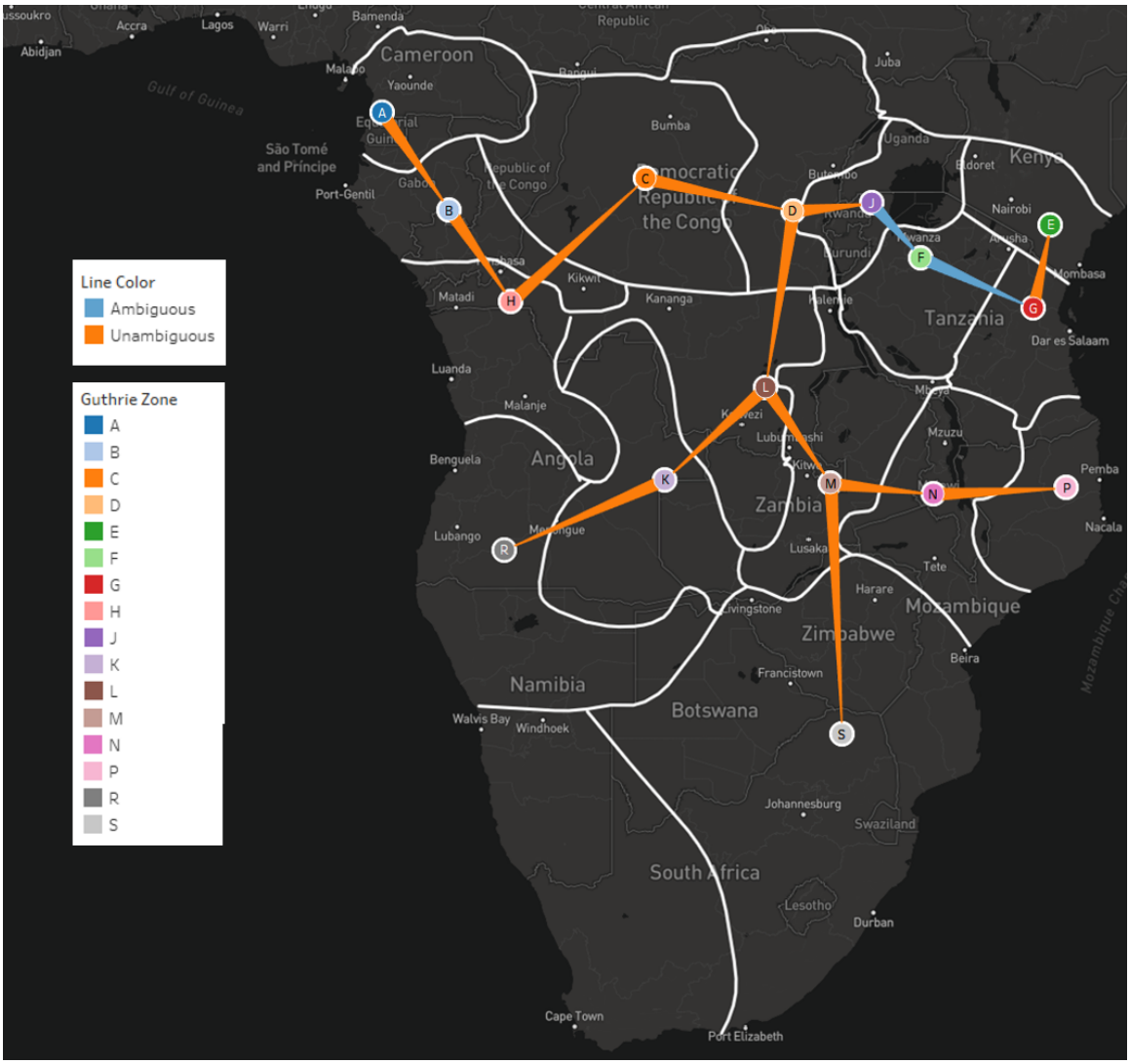

Figure 9: Migratory model generated using the Guthrie zone transitions from the combined data and the haversine distance-based Sankoff matrix.

The migratory model generated using only mtDNA data shows a slightly different trajectory. The path starts in the northwest in Guthrie zone A and moves southward to zone B, but seems have an area of ambiguity between zones B, C, H, and D. At zone D a split occurs with migration going east through zones J, F, G, and E and south to zone L, similar to the combined model. From zone L, another split occurs, heading west to zones K and R and heading southwest to zone M. At zone M, another split occurs with one branch heading south and ending at zone S and heading east to zone N and ending at P, the same as the combined model. See figure 10. Note that there are more ambiguous transitions occurring here attached to zones (B, C, and H), (J, E, F, and G), and (D, L, M, and N).

The model generated using only the Ychr data is strikingly similar to the combined model. This models shows the same trajectory starting in the northwest in Guthrie zone A and moving southward to zone B and H and then an eastward movement through zones C and D. At zone D a split occurs with migration going east through zones J, F, G, and E. The only topological difference in this model compared to the combined is that the southbound movement from zone D is linked to zone M rather than H, but this is an ambiguous transition. At zone M, the branching is the same to the combined model in that one is heading south and ending at zone S and one is heading east to zone N and ending at P. See figure 11. Note that a major difference in ambiguity from the combined model is along the path through H, C, D, J, and G.

The final separate model to compare is the cultural model. The model generated using only the cultural data is strikingly similar to both the combined model as well as the Ychr model. The trajectory is seen starting in the northwest in Guthrie zone

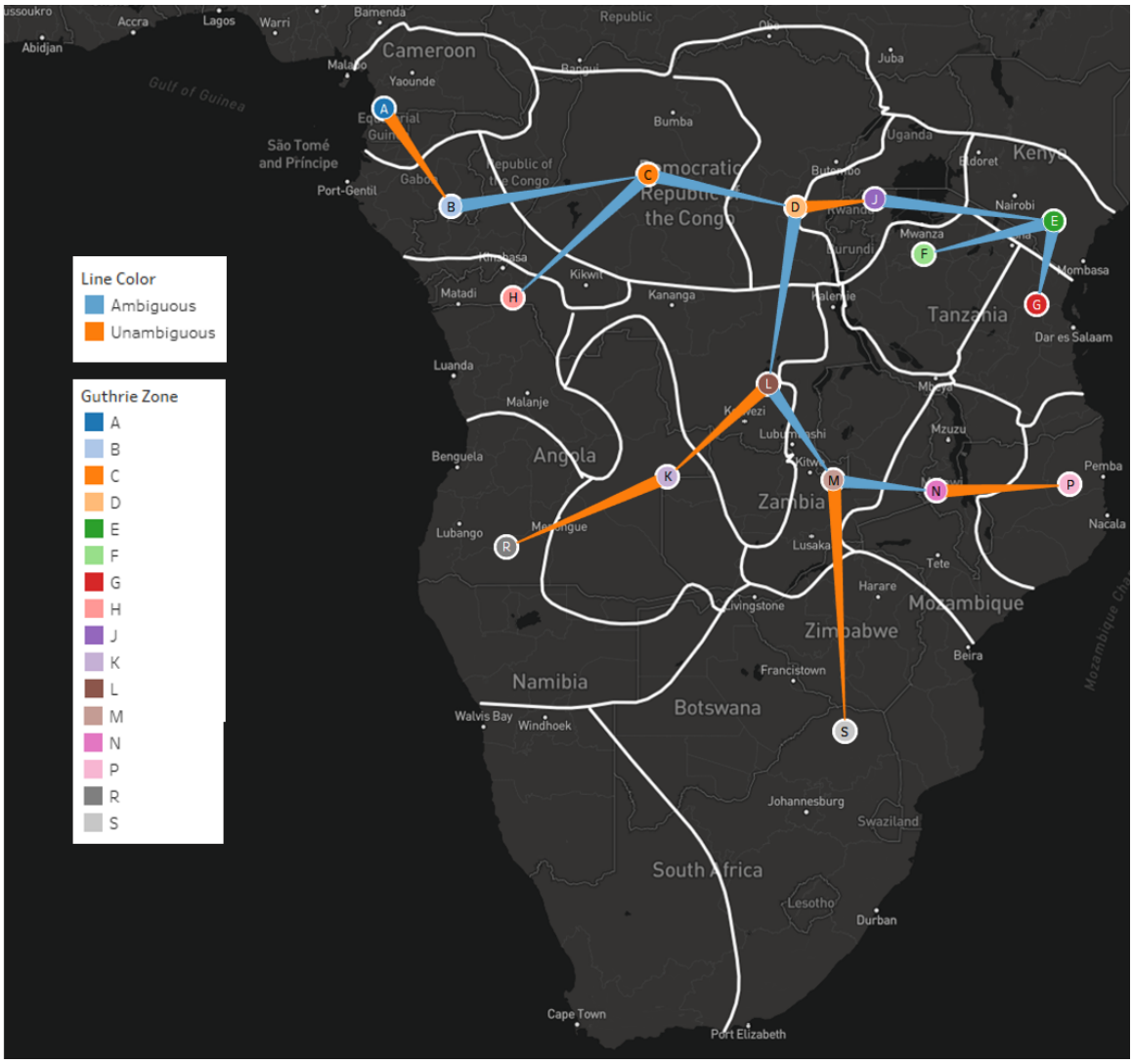

Figure 10: Migratory model generated using the Guthrie zone transitions from the mtDNA data and the haversine distance-based Sankoff matrix.

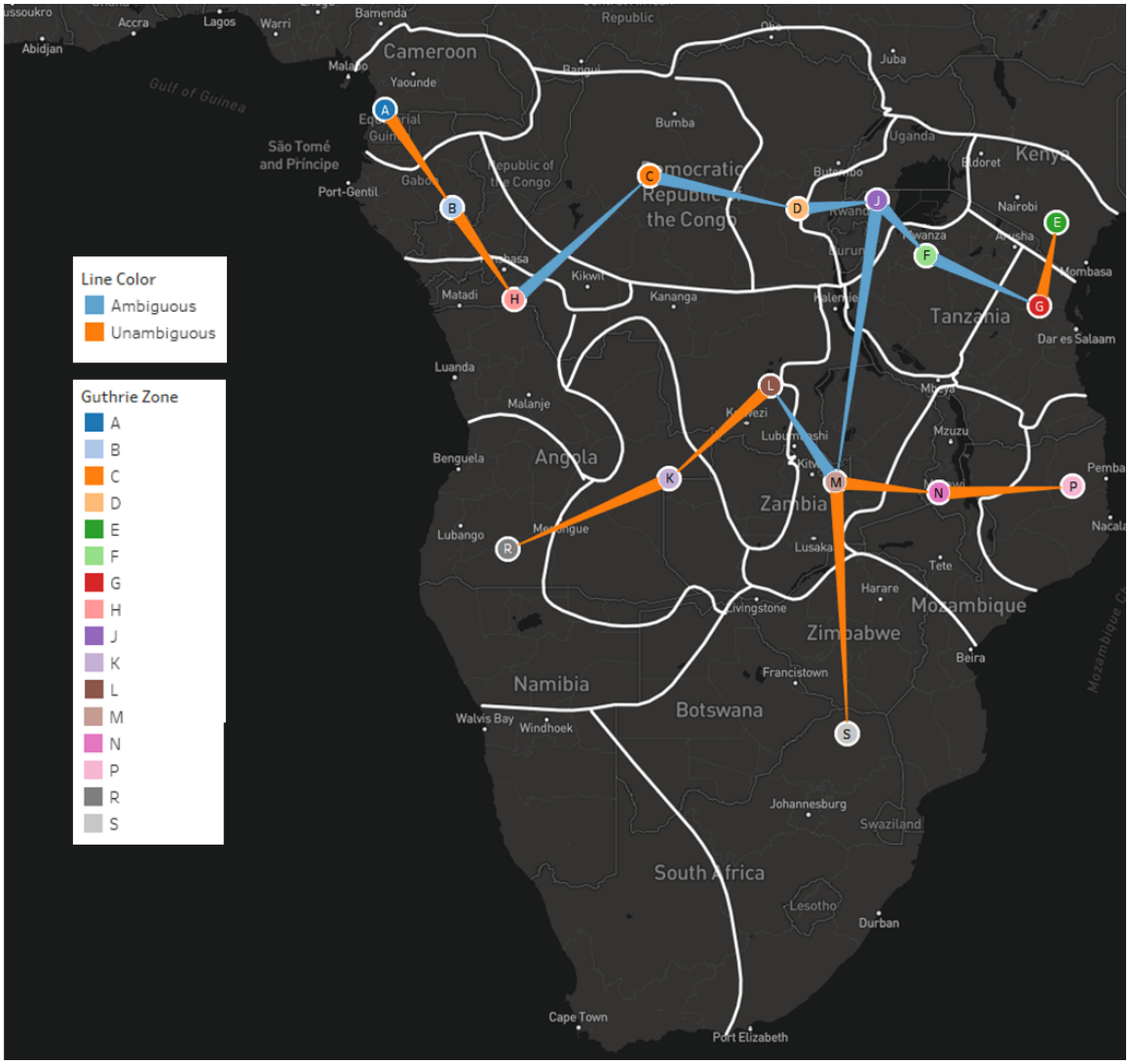

Figure 11: Migratory model generated using the Guthrie zone transitions from the Ychr data and the haversine distance-based Sankoff matrix.

A and moving southward to zones B and H. Then, the model shows an eastward movement through zones C and D. At zone D a split occurs with migration going east through zones J, F, G, and E. The major difference in this model compared to the combined is that the southbound movement from zone D is linked to zone K rather than H, but this is an ambiguous transition. At zone M, the branching is the same to the combined model in that one is heading south and ending at zone S and one is heading east to zone N and ending at P. See figure 12. Note that the only ambiguous areas of this model are mainly in the center of the continent, which is the area of disagreement to the combined model.

2.4 Discussion

Overall, the trees have similar overall topologies. This is seen in the reconstructed migratory paths as well. For example, there is an overall agreement that there is a lateral movement from Guthrie zone A to the west. Also, around zone D, there seems to have been a split in lineages with a new trajectory moving south. Overall, there is an agreement in the branching pattern seen in the southern zones as well. However, the points of disagreement are often in the central zones on the region. The tree topologies all differ slightly, but the common monophyly of Guthrie characters F, R, P, and S among all the results from different datasets suggests that the endpoints of the Bantu migration have been largely stable over time with respect to the genetic and cultural makeup of the population.

Comparing the migratory models, the combined model has the clearest migration path as it has the least ambiguous transitions and a more reasonable trajectory. For

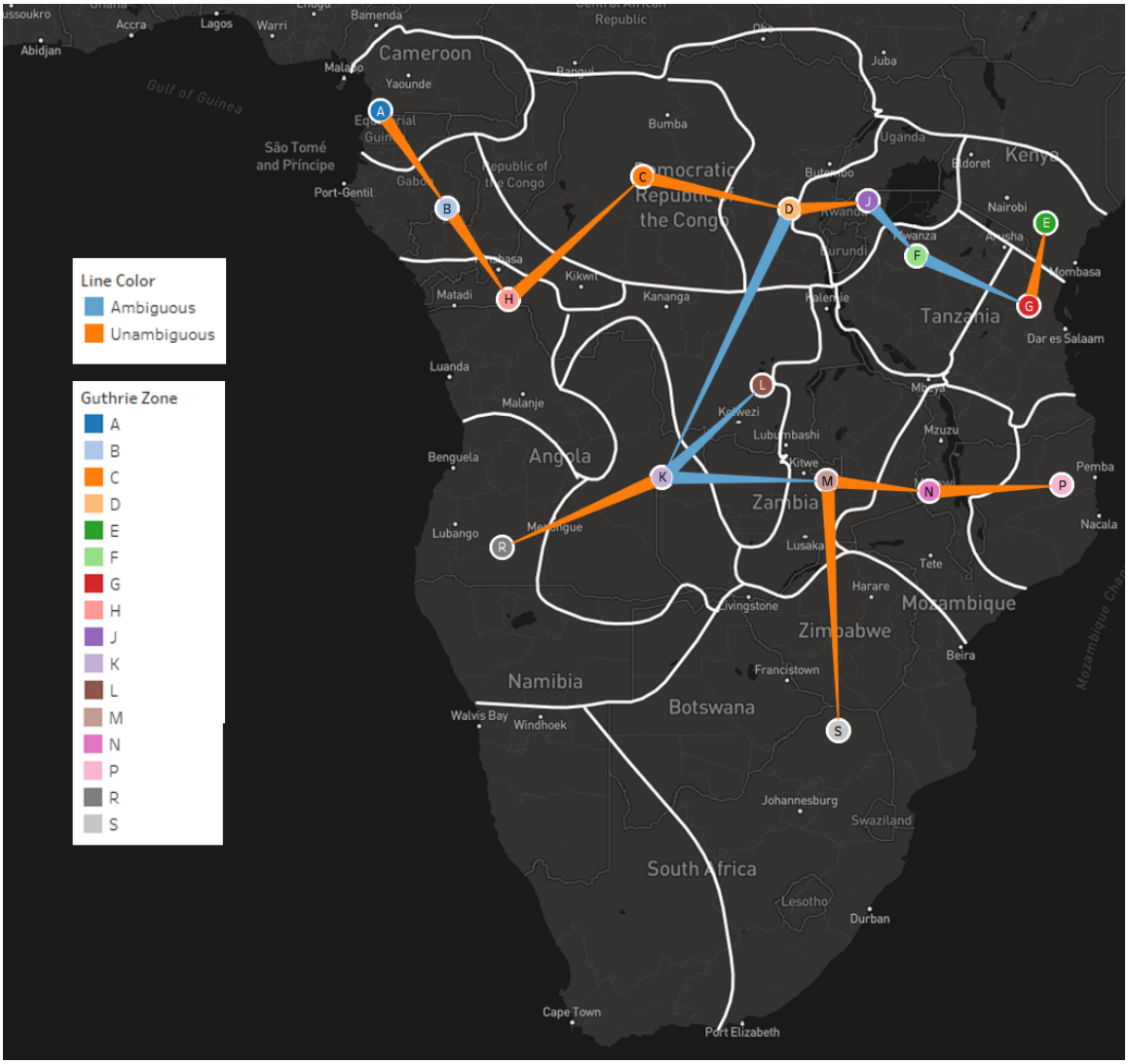

Figure 12: Migratory model generated using the Guthrie zone transitions from the cultural data and the haversine distance-based Sankoff matrix.

example, in comparison to the mtDNA migratory model, the combined model exhibits far less "back tracking" across the continent. In the mtDNA model, there are some transitions, such as from zones J to E and back to F, that are not seen in the combined model. It seems the movement in the combined model is mostly unidirectional until branching begins to occur. Thus, there are novel results here that have very little internal conflict and become less ambiguous when data are combined.

By combining the datasets together, a more cohesive migration postulation is created. This combined model takes aspects from each of the constituent sets, and traces the trajectory in a more condensed manner. This can be seen in figure 13. Note the combined model (the blue path) is central to the other individual models' paths.

Finally, both the combined migratory model, as well as the separate models, provide evidence that partially confirms the initial hypothesis that the migration is characterized by an "early split". This can clearly be seen in Guthrie zone D in the combined model, as well as in most of the separate models. These pathways are different than previous authors have reconstructed. Specifically, the location of the split is different (more centered on the continent) than the "early split" de Filippo model. In addition, there is evidence of later, more branching migratory paths after the primary split. The branching events can be seen in the models in zones L and M, but with less ambiguity in the combined model.

A southeastward lineage movement is seen from zones A to B to H to C. This path covers from modern-day Cameroon to Equatorial Guinea and Gabon, to the Republic of the Congo and then to the Democratic Republic of the Congo (DRC). Next, there is a split at zone D. One lineage heads eastward from zones D to J to F to G to E.

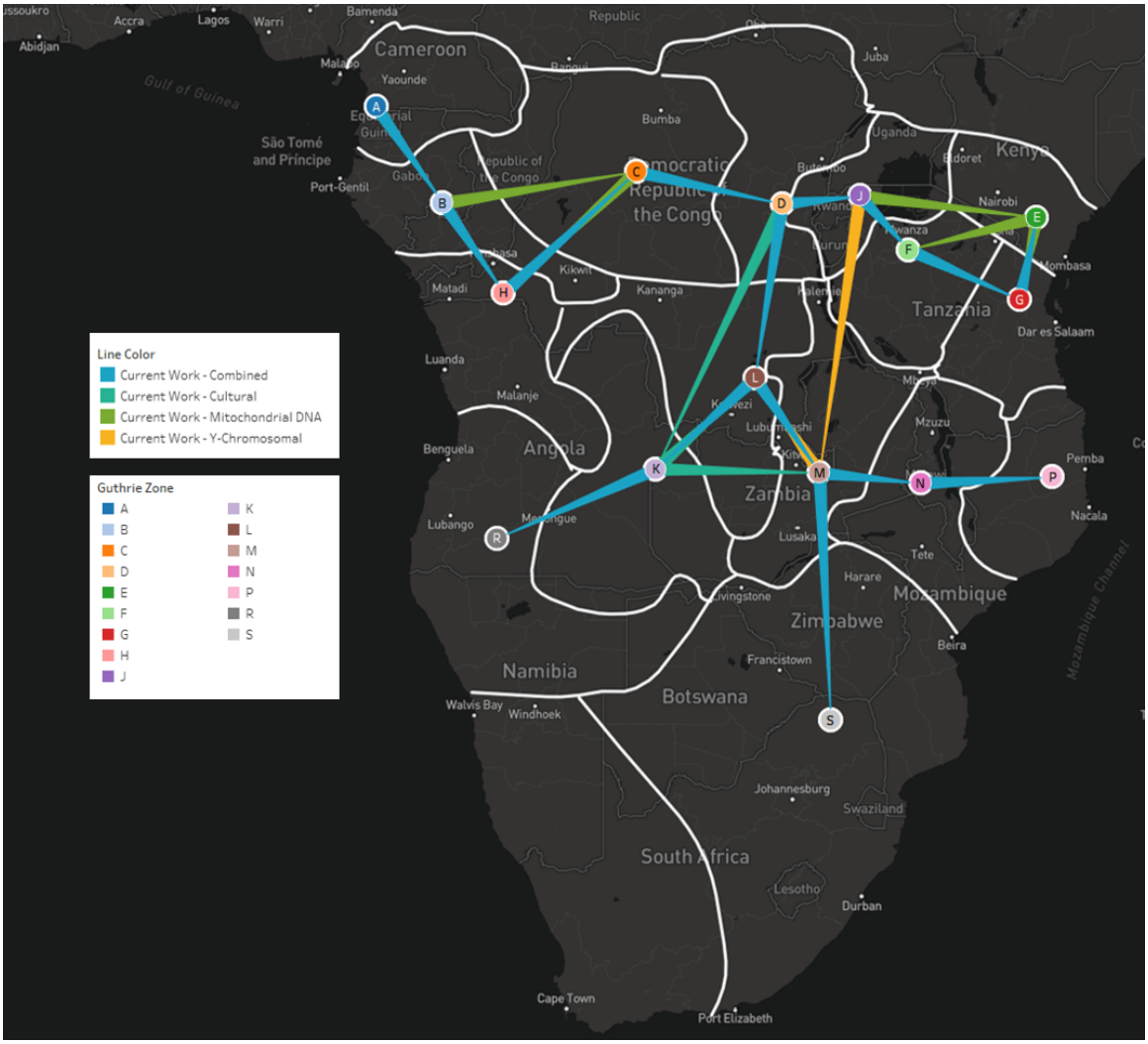

Figure 13: All four generated migratory models overlaid for comparison.

This path covers from modern-day DRC to Uganda to Tanzania and then to Kenya. The ancestral state reconstruction of zones J to F to G to E are the only ambiguous ones on the map.

Another lineage that splits from zone D runs southward to L in the southern part of modern-day DRC. From zone L, one lineage splits eastward to zones K and R in modern-day Zambia and Angola. Also, from zone L, another lineage splits westward to zone M in modern-day Zambia. From zone M, there is another split with one lineage moving due east through zones N and P, which represent modern-day Malawi and Mozambique. Lastly, from zone M, there is a lineage that moves due south to zone S, which is modern-day South Africa.

Overall, the approach with combining the datasets together has provided a more confident migratory model for Bantu migration than that of any of the separate datasets. By looking at paternal, maternal, and cultural data together, a clearer trajectory is drawn with far less ambiguity in the overall path.

CHAPTER 3: COMPARING TAXA PROXIMITIES USING DIMENSIONALITY REDUCTION

3.1 Introduction

In addition to phylogenetic tree generation, further analysis of the data is performed using two dimensionality reduction techniques. Phylogenetic trees use inference methods that rely on an optimality criterion such as maximum likelihood or maximum parsimony. Along with the parsimonious inference method, Multidimensional Scaling and Laplacian Eigenmaps are used as distance-based methods for comparison.

In the previous chapter, the outcome of the parsimonious tree generation shows that the combined data produces a less ambiguous model of the Bantu migration. Now, using dimensionality reduction, this model's data is further analyzed to understand the relationships or distances between taxa by projecting into a lower dimensional space. This helps to identify taxa that are closely related in the data, but are not in the same Guthrie zone. Then, rolling up to the Guthrie zone-level, the trajectory is compared to the placement of points in the lower dimensional space.

Agreement between the outputs of the tree generation, probabilistic simulations, and the dimensionality reduction exercises are used to build support or confidence in the Bantu migration model. Any non-negligible differences between the results help to pinpoint the areas where confidence is low in the migration model.

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

Research performed by Verdu et al. [53] around the genetic and linguistic trajectories of Creole people served as a model for the analysis of the Bantu data. The research was completed using a multidimensional scaling analysis to look at individual-pairwise allele-sharing dissimilarities of different populations. Also, the research shows the use of the k-Means clustering method to analyze admixture among the groups. From a visualization standpoint, the paper includes graphics generated from the results of the dimensionality reduction steps that were useful in the overall assessment of variation in Creole people.

3.2 Materials and Methods

3.2.1 Data Shaping

Given that the available data consists of sequences in different shapes and from different file formats, the data must be transformed such that the desired algorithms can use it. Generally, the required shape is rectangular or tabular in that every row is an observation and every column is a feature. Specifically, the data is transformed such that each row of the data is an individual sample (by taxa) and then each column is an attribute or sample of taxa (Guthrie zone, TaxaID, etc. plus mtDNA, Ychr, and Cultural sequences) each as its own column.

The individual data files from the mtDNA, Ychr, and Cultural datasets are each loaded into their own SQL database table. In total, there are four tables being used: one table for each of the three datasets and one table to house the geographical

Table 6: Schema of the mtDNA database table.

mtDNA	Example
TaxaID	H31Yaka
GuthrieZone	H
Taxa	H31Yaka_BiakaPygmy_NA_mt01
BCMID	H31
SampleID	mt01
mtDNA	gatcacaggtctatcacccta

Table 7: Schema of the Ychr database table.

Ychr	Example
TaxaID	A34Benga
GuthrieZone	A
Taxa	A34Benga_Benga_Gabon_chry01
BCMID	A34
SampleID	chry01
Ychr_STR	EJGJBPDBFEBB

information for mapping. See figures 6, 7, 8, and 9. See appendix A for the SQL code used to generate these tables.

TaxaID is used as the key on which to join the tables together to make a single file. Two views are created: one inner joining the data and one using a full outer join. This created two versions of a combined dataset for use in the dimensionality reduction exercises. The inner join view uses only perfectly matched TaxaIDs between all three datasets. In contrast, the outer join data joins on TaxaID, but can leave a particular dataset's column as NULL if there is no match in the data.

Table 8: Schema of the Cultural database table.

Cultural	Example
TaxaID	A242Duala
GuthrieZone	A
Taxa	A242Duala
BCMID	A24 2
Cultural_EthnogAtlas	0101818559189B293919997

Table 9: Schema of the Geographic Information database table.

Geographic Information	**Example**
GuthrieZone	B
BCM ID_FULL	B80L1
BCM ID_SIMPLE	B80
BCM ID_SUFFIX	L1
BCM ID_INDEX	1
BCM NAME	Lwel 1
BCM FULL NAME	B80L1Lwel 1
Longitude	-4.45
Latitude	20.1

An R script was written and used to further prepare this data for use in the dimensionality reduction algorithms. Since the data consists of strings of sequences, conversion has to occur to get a numerical representation of the characters in each position (nucleotides for the mtDNA data, character state labels for alleles in the Ychr data, and answers to the cultural questions in the Cultural data).

To begin, the data is loaded into R from the joined SQL view. At this stage, the values are character strings of the entire sequence from each data source. See step (1) in figure 14.

The strings are then split up and each position is assigned to a new variable. This causes the data table to grow considerably wider as the length of the sequences are quite long. See step (2) in figure 14.

At this stage, the data is still solely character-based. So, the next step is to dummy code these values to convert them from single character strings to 1s and 0s [49]. This process increases the number of variables further by creating a new column for each possible character at that position. For example, if position 145 of the mtDNA sequence can have “A”, “C”, “T”, and “G”, four dummy columns are

created to capture these levels. The data at this step is quite large as it now has >30,000 variables and 22,000-33,000 rows. See step (3) in figure 14.

Next, the data is rolled up to either the TaxaID or Guthrie zone level. (The dimensionality reduction is performed and visualized at both levels for comparison.) By summarizing and returning the mean value for each column, this now captures the percentage that a character occurs for each Taxa or zone. For example, "44% of the time, individuals from Guthrie zone A have "g" in the first position of their mtDNA sequence." Each set of columns for a position will sum to 1. See step (4) in figure 14.

At this stage, the data now holds a numerical representation for each position of the sequences for each TaxaID or Guthrie zone. Now the data is in an acceptable format for the dimensionality reduction algorithms. The packages that were used to read in, shape, and dummy code data in R are: *tidyr* [56], *readr* [57], *dplyr* [55], *purrr* [28], and *psych* [45].

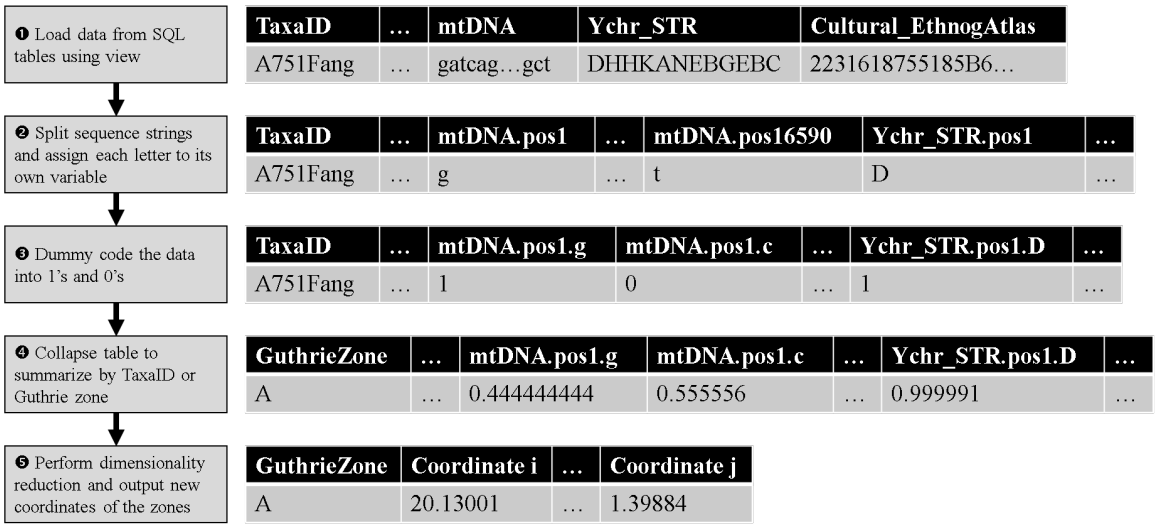

Figure 14: Dimensionality Reduction data shaping workflow.

This dimensionality reduction pipeline was completed using an R script. The R script includes sections to load in the data, perform the aforementioned shaping, and output the Multidimensional Scaling and Laplacian Eigenmap analyses, then writes the results back to disk. See appendix B.

The 2-dimensional result sets are visualized via scatter plots using Tableau Desktop, which allowed for interactive analysis of the data [48].

3.2.2 Multidimensional Scaling

Multidimensional Scaling (MDS) is a technique often used to reduce dimensionality [35]. 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*. This analysis can be completed 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 were calculated as this is the limit to what can be visualized.

The steps of the classical MDS algorithm are as follows:

1. Begin by calculating the Euclidean 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.

The non-metric version of multidimensional scaling finds the non-parametric monotonic relationship, as well as the Euclidean distances between the distance matrix entries [5]. Non-metric multidimensional scaling differs from classical MDS in that the algorithm will iteratively sample the data, scale it, and then calculate the stress.

The steps of the non-metric MDS algorithm are as follows:

1. Sample from a normal distribution to collect a random arrangement of the datapoints.

2. Recalculate the distance matrix given the arrangement.

3. Determine the optimal monotonic transformation function of the distances to result in optimally-scaled data.

4. Compute the stress value of the scaled data.

5. If the stress value is below some threshold or if the algorithm has reached convergence, exit the algorithm. Otherwise, return to step 1.

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

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 num-

ber of dimensions for the analysis of the data. The result of an MDS analysis is a 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).

Non-metric MDS is sensitive to certain parameter inputs such as the maximum number of iterations, the power for the Minkowski distance of the configuration space, and the tolerance for declaring convergence [36]. The algorithm is $O(n^3)$, so the manipulation of these parameters can provide varying time performance of the analysis. Since the data here have a small number of observations, the convergence tolerance can be small and maximum number of iterations can be quite large without significantly sacrificing speed.

Non-metric MDS is available as an R function in the *MASS* package called *isoMDS* [52].

3.2.3 Laplacian Eigenmaps

As a sub-comparison to MDS and, by extension, the phylogenetic tree result, Laplacian Eigenmaps (LEs) were also used. While MDS may provide an acceptable projection of the data in a lower-dimensional space, 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.

LE's look for local similarities only. This is accomplished by defining a neighborhood using the k-Nearest Neighbor (k-NN) algorithm, the epsilon neighborhood algorithm, or some other distance matrix method. Using the Gaussian kernel, Eu-

clidean distances are converted to similarity values [3]. The desired number of reduced dimensions must be defined *a priori*. The LE analysis is run using 2 and 3 dimensions given that this is the limit to what can be visualized.

The steps of the LE algorithm are as follows:

1. Construct the adjacency graph using either the k-NN algorithm or the epsilon neighborhood algorithm.

2. Choose the weights of the edges of the resulting network either using a heat kernel or a simple-minded (non-parameterized) approach.

3. Calculate the eigenmaps of the graph and output the eigenvectors to embed the data into an n-dimesional Euclidean space.

Given that the output of the LE algorithm is highly dependent on the graph generation, the optimal parameters must be found for the chosen graph algorithm. It was opted for k-NN to be used as the graph algorithm for this analyses. Before the LE algorithm was run, the optimal number for k in the k-NN was determined based on a parameter sweep for values of k between 1 and the number of rows in the data. (The number of rows is equivalent to the number of TaxaIDs or Guthrie zones in the input data, depending on the dataset being analyzed.) Then, the value for k that resulted in the highest accuracy for the generated k-NN model was selected and passed to the LE algorithm. This ensures that the LE algorithm is using the k-NN graph with the highest possible accuracy given the input data.

These steps were completed for both inner and outer join datasets, TaxaID and Guthrie zone summarizations for 2 and 3 dimensions. This results in eight datasets

that are then visually analyzed for comparison to the phylogenetic results. See table 11 for the accuracy levels and optimal k values for each input dataset.

LE is implemented as an R package called *dimRed* [34].

3.3 Results

Using the four generated datasets of summarized data (using the inner joined and outer joined data and summarized by Guthrie zone and TaxaID), the dimensionality reduction analyses are completed. This results in individual data outputs ready for visualization. Note that the inner joined data results in a lower number of samples and thus a lower coverage of Guthrie zones, but still generally covers most regions of the continent.

Using Tableau Desktop, visuals for the dimensionality reduction results are created with the 2-dimensional datasets (both inner joined and outer joined data). This includes 4 LE visuals and 4 MDS visuals as well as a map plotting the location of each of the TaxaIDs.

The results of the dimensionality reduction analyses require a more involved approach than simply looking at the plots at face value. For MDS and LE, the locations of the points are indicative of the similarity of the TaxaID or Guthrie zone. However, some of the plots have densely packed areas that are difficult to see into without further investigation. To do so, the plots are linked to the TaxaID Map (see figure 29) so that selecting points on a particular plot filters the map to show those points as well. This interactivity allows for geospatial comparison of the points to their location in the dimensionality reduction plots.

In the 2D MDS plot using inner joined data summarized by TaxaID (see figure 15), most TaxaIDs are densely located in the lower left corner with the two points (S33Sotho and M52Lala) from the S and M zones located farther away in different quadrants. Zooming in on the dense area shows a much clearer distribution of the TaxaIDs. See figure 18. The location of the TaxaIDs with regard to their corresponding Guthrie zones is overall very similar to the results seen in the combined migratory model. For example, it appears that Guthrie zones A and B are similar. Then, zones K, L, and M in the middle and R, N, and S are also at the tips of the migratory path.

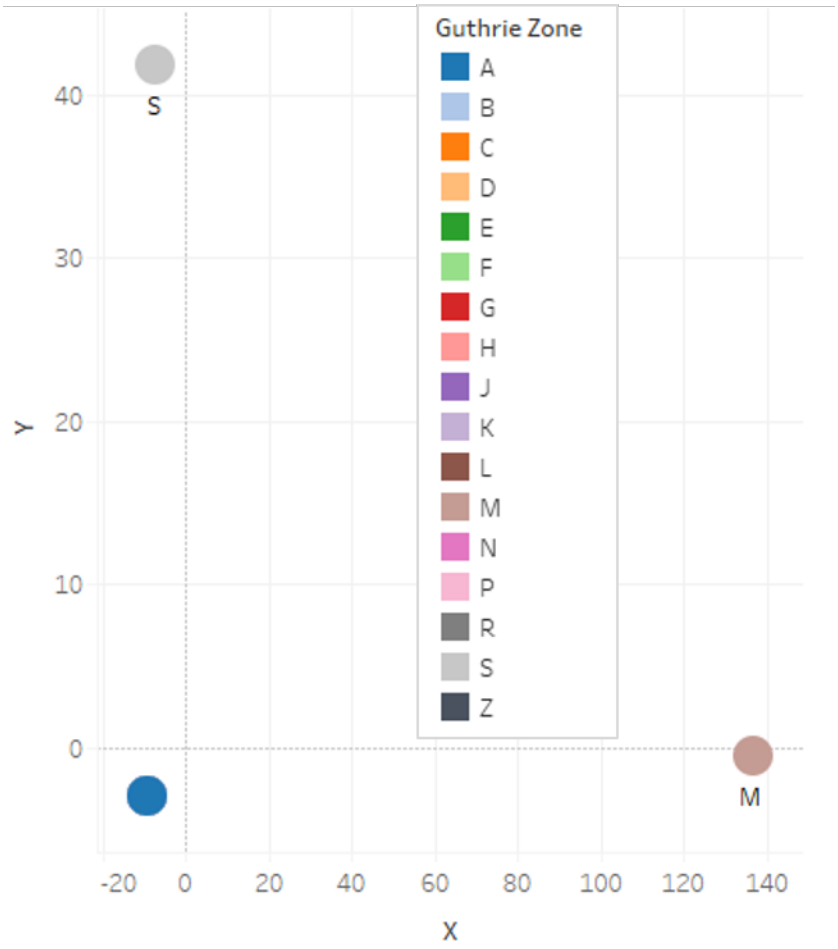

Figure 15: 2D non-metric MDS scatter plot of inner joined data summarized by TaxaID.

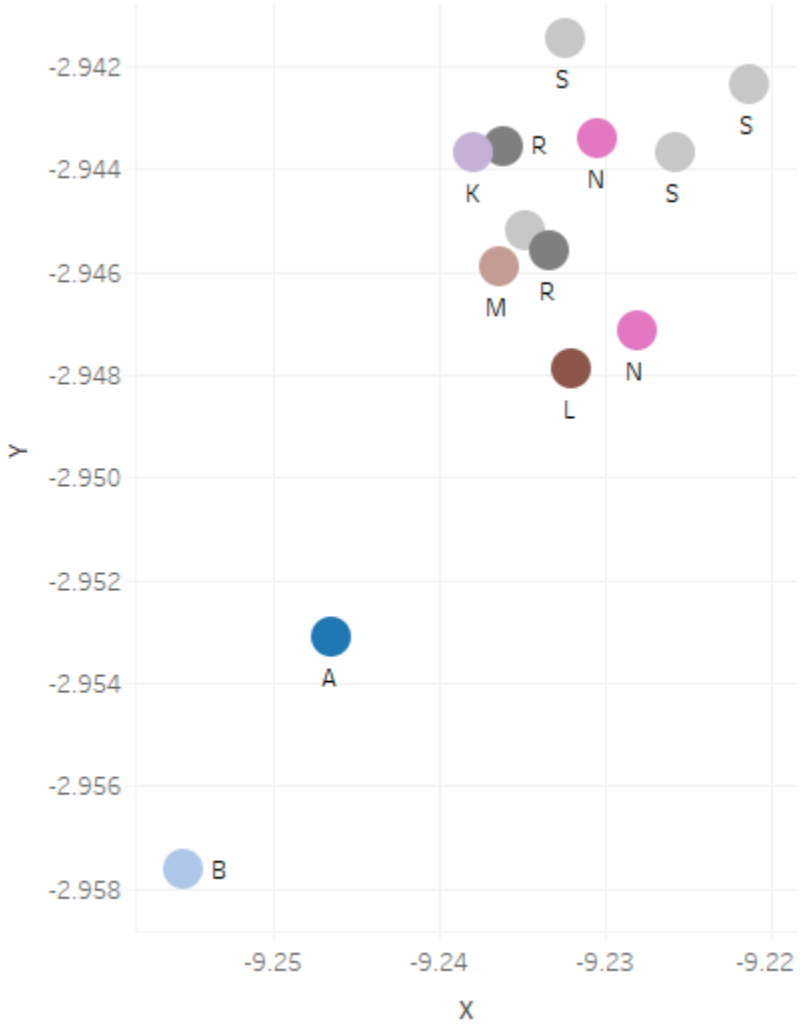

Figure 16: A zoomed-in view of figure 15, disregarding the farthest two points (S33Sotho and M52Lala).

Next, in the 2D MDS plot using the outer joined data summarized by TaxaID (see figure 17), the image is quite similar to before. Most of the points are located in a single cluster with a few TaxaIDs placed farter away. This time, the 11 points from the A, B, F, H, K, L, and S zones are located farther away. Zooming in on the densest area shows a less clear distribution of the TaxaIDs. See figure 18. The location of the TaxaIDs with regard to their corresponding Guthrie zones is partially discordant to the results seen in the combined migratory model. However, it does appear that the placement of most of the TaxaIDs with respect to their Guthrie zones is acceptable. For example, points from the more southern zones K, L, M, R, and S are closely related, which is likely true. The northern zone seems to be spread around the southern zone cluster in the center.

After looking at the MDS plots that were generated used data summarized by TaxaID, a secondary look is taken at the results using data that was generated by summarizing over Guthrie zones. These datasets are smaller as the maximum number of records is 16 (or 17 if the outgroup zone Z is included).

First, using the inner joined data, another MDS analyses is run and plotted, resulting in figure 19. Notice that the M Guthrie zone is shown far from the seemingly vertical line of the remaining zone. However, excluding zone M results in a much more interesting spread of the other zones. See figure 20. Moving from left to right, the location of the zones is similar to the combined migratory model. For example, zones A and B are in the north and then the remaining zones are after the "early split" thought to have occurred in zone D.

Finally, the last MDS analysis is run using the outer joined data. The outer join

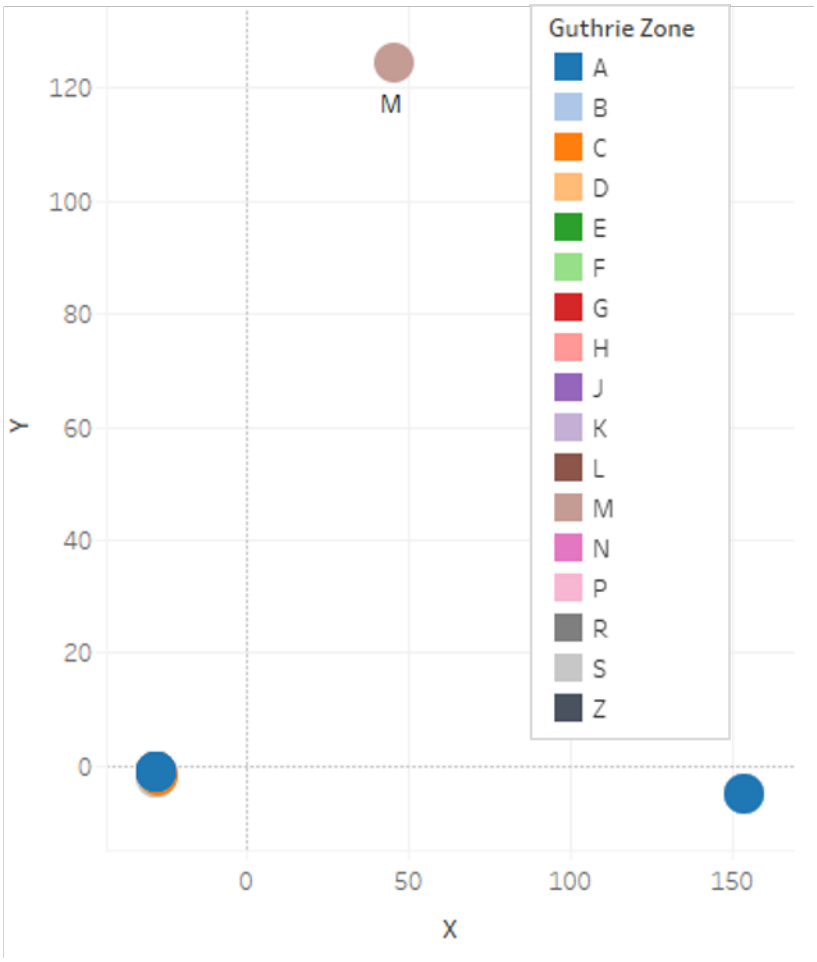

Figure 17: 2D non-metric MDS scatter plot of outer joined data summarized by TaxaID.

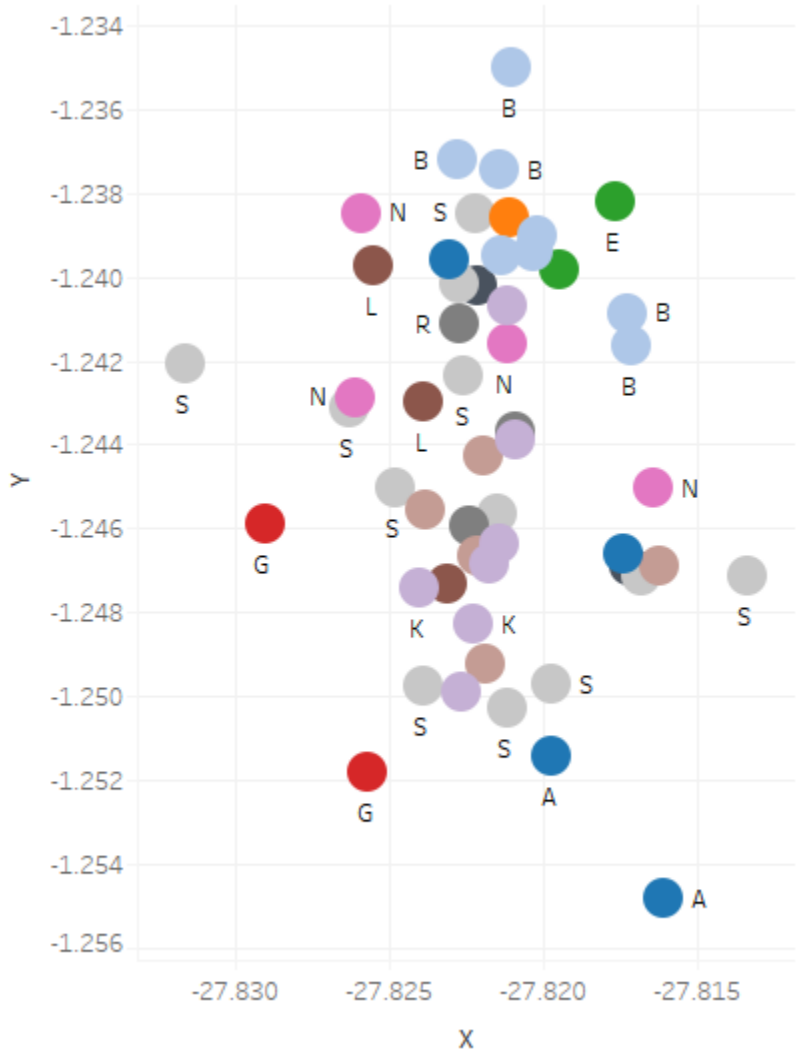

Figure 18: A zoomed-in view of figure 17 of the densest cluster of points.

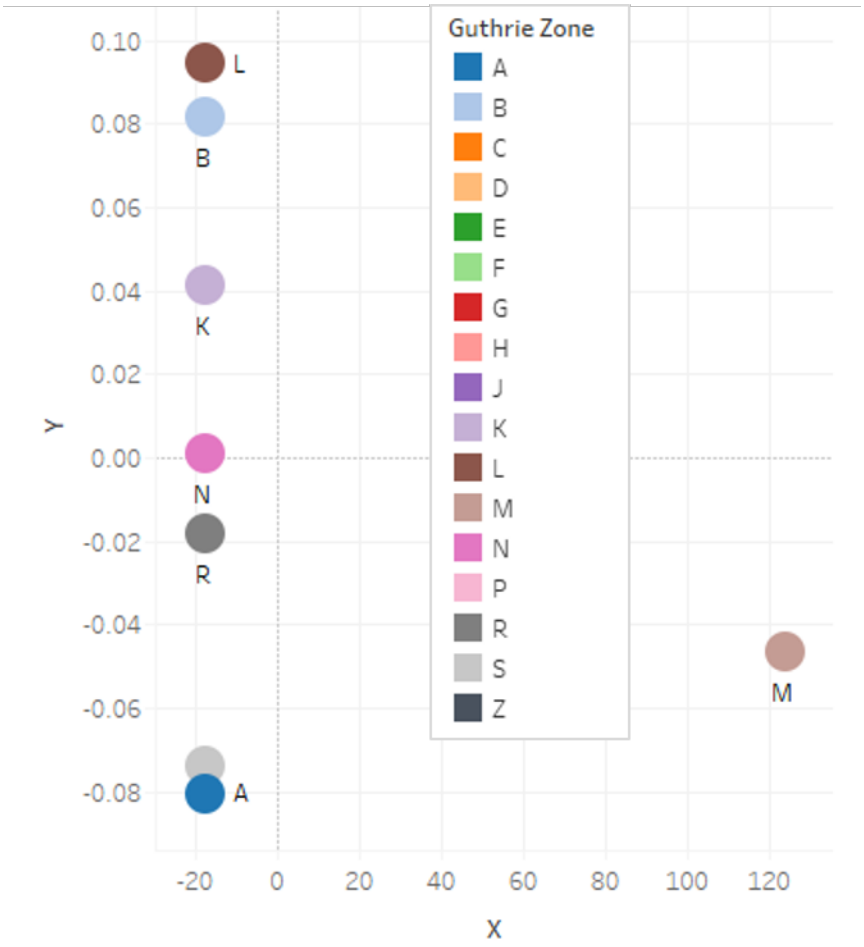

Figure 19: 2D non-metric MDS scatter plot of inner joined data summarized by Guthrie zone.

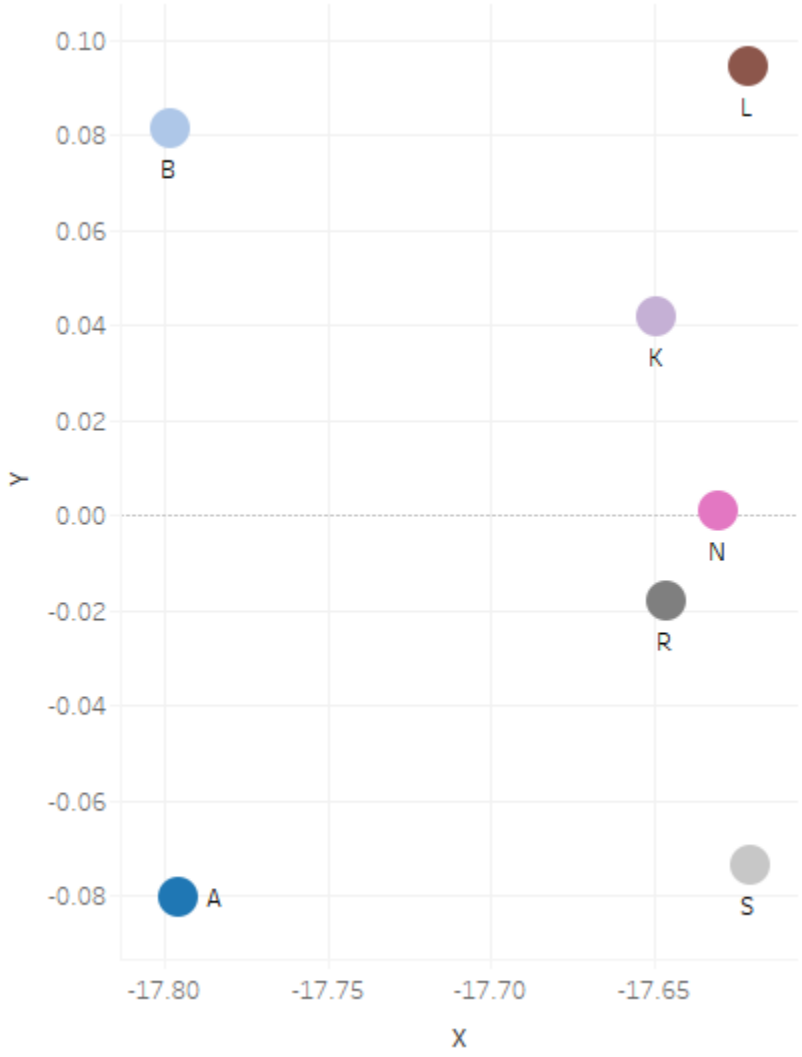

Figure 20: A filtered view of figure 19, disregarding Guthrie zone M.

data provides coverage for almost all the Guthrie zones. In figure 21, the spread of the zones is obfuscated due to zones M, F, and A. Figure 22 shows the plot after excluding these points. This analysis, in particular, is the most discordant to the combined migratory model. There seems to be no north-to-south or east-to-west flow between the points. Even geospatially-close zones are not in an expected proximity to one another. For example, zone G (from the northeast) is very close to zone R (from the mid-southeast), but not close to zone C, to which it is adjacent geospatially.

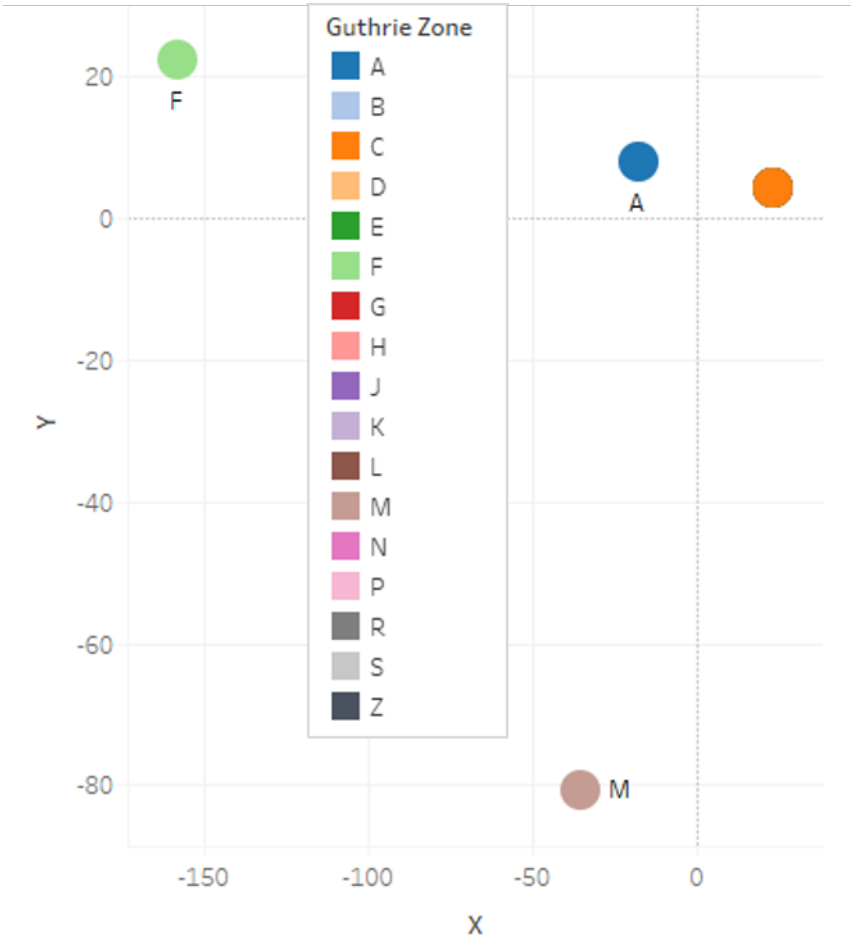

Figure 21: 2D non-metric MDS scatter plot of outer joined data summarized by Guthrie zone.

Transitioning to LE analyses, more extreme cases of dispersion than in MDS were

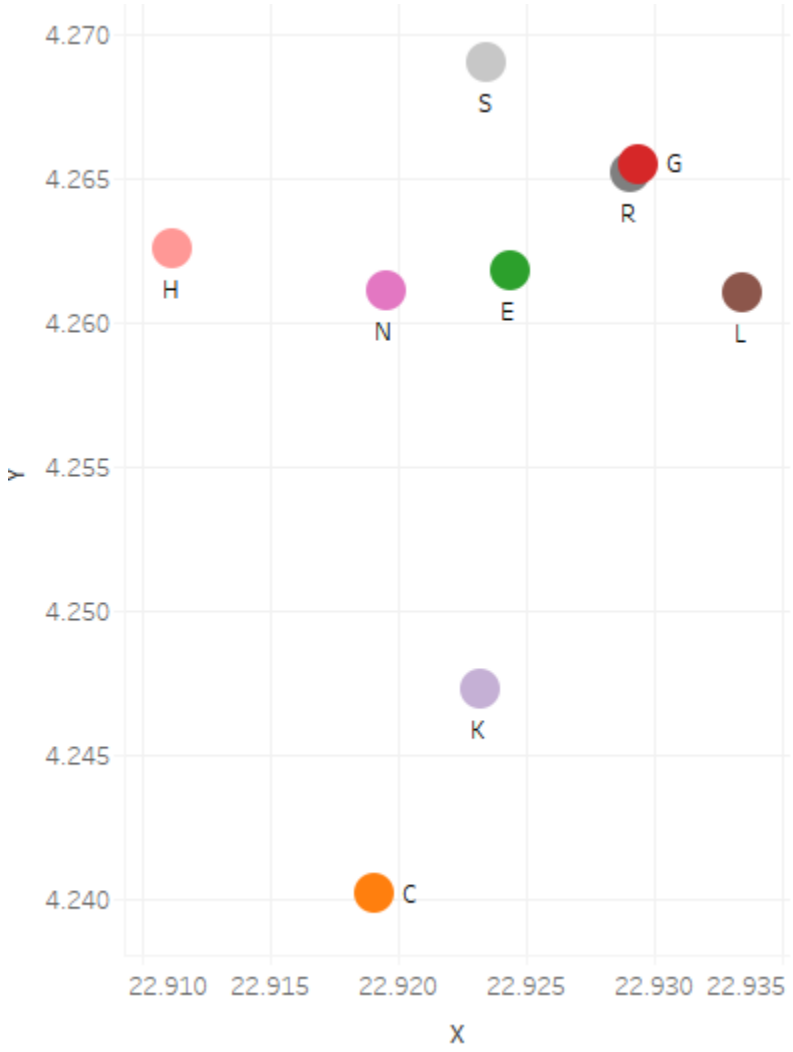

Figure 22: A filtered view of figure 21, disregarding Guthrie zones A, F, and M.

expected. However, this is not necessarily the case for all of the following plots.

For the 2D LE plot using the inner joined data summarized by TaxaID, the locations of the points are well-contained in a dense cluster except for two TaxaIDs (K11Ciokwe and R111Umbundu) belonging to the K and R Guthrie zones, respectively. Zooming in on the denser cluster shows a radiating pattern of the similar points compared to the combined migratory model. See figure 24. Specifically, if points belonging to zones A and B are taken as a center, the points moving outward are all south. Also, if the TaxaID belonging to zone L is taken as a starting point, movement to N, S, and R is consistent with the combined migratory model.

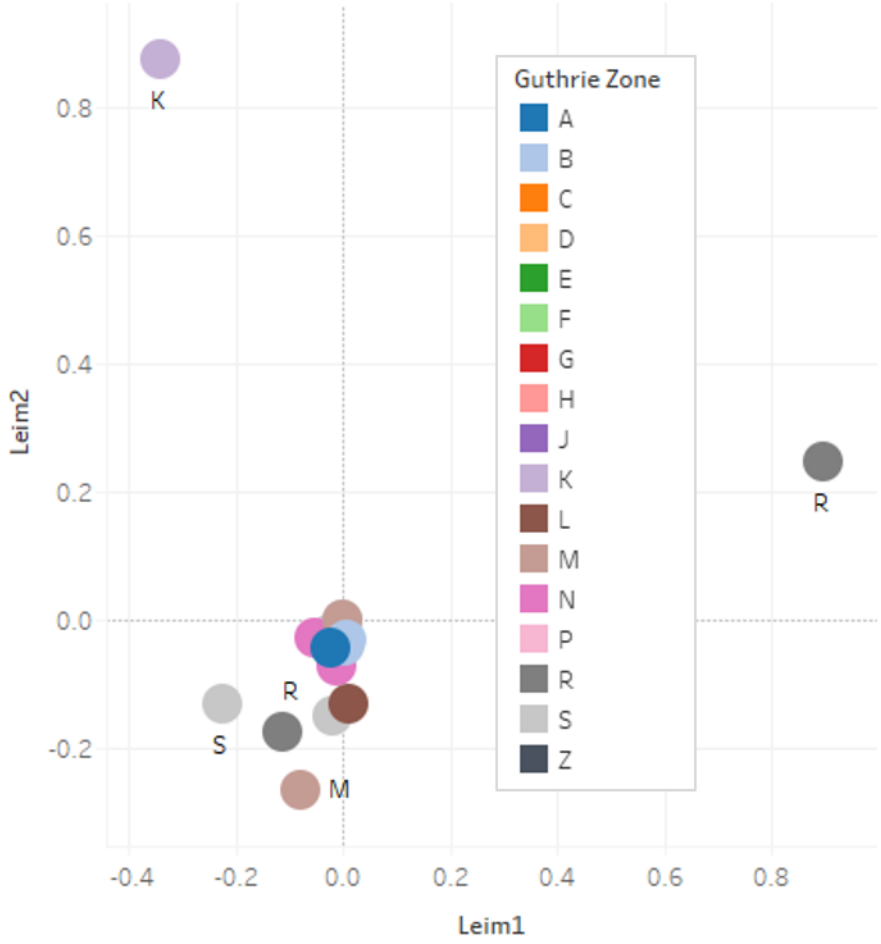

Figure 23: 2D LE scatter plot of inner joined data summarized by TaxaID.

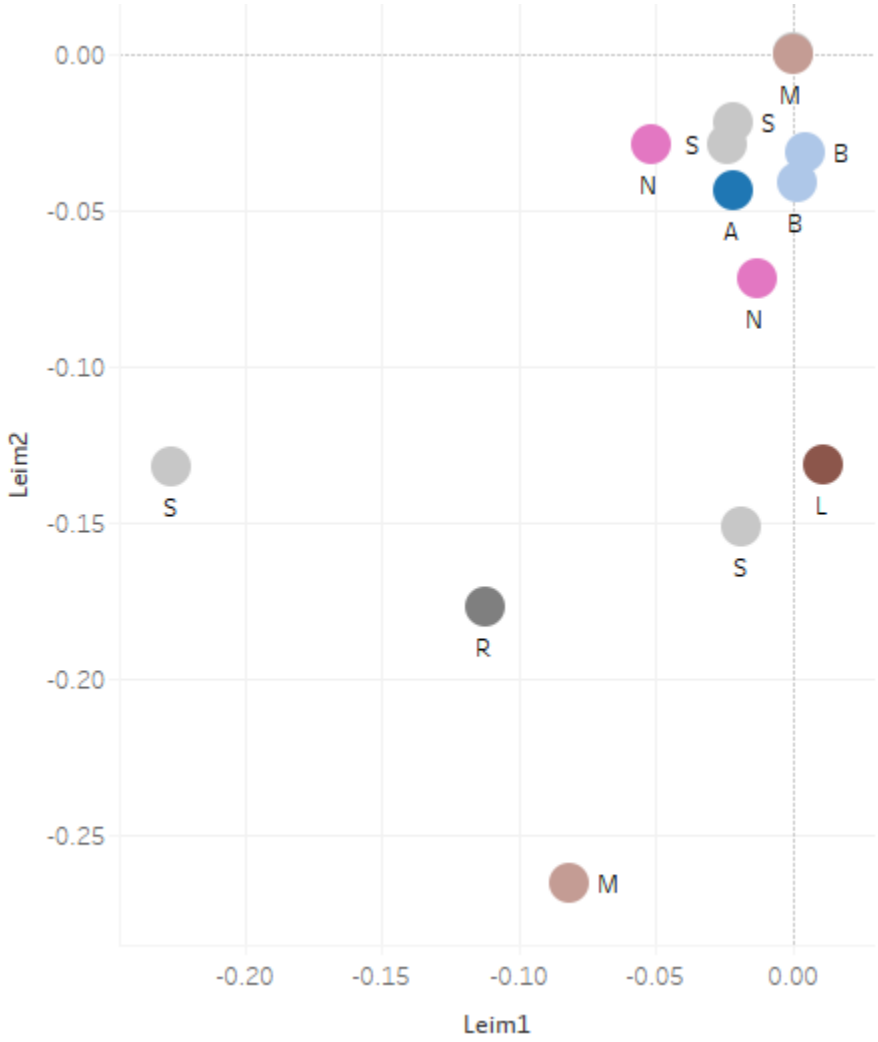

Figure 24: A zoomed-in view of figure 23, disregarding the farthest two points (K11Ciokwe and R111Umbundu).

Next, using the outer joined dataset summarized by TaxaID yields a plot very similar to the MDS plot in figure 19. See figure 25. Two TaxaIDs in Guthrie zones C and M (C55Kele and M54Lamba) appear to be skewing the locations of the rest of the zones by making them appear as if they are in a vertical line. Excluding these results in a slightly clearer picture. It appears that there is some, but not complete, agreement to the combined migratory model. For example, there is a cluster depicting a spread of northern to middle zones, such as B to H and then to M and P, but it is

difficult to explain the close placement of TaxaIDs in zones A and S, or G and N.

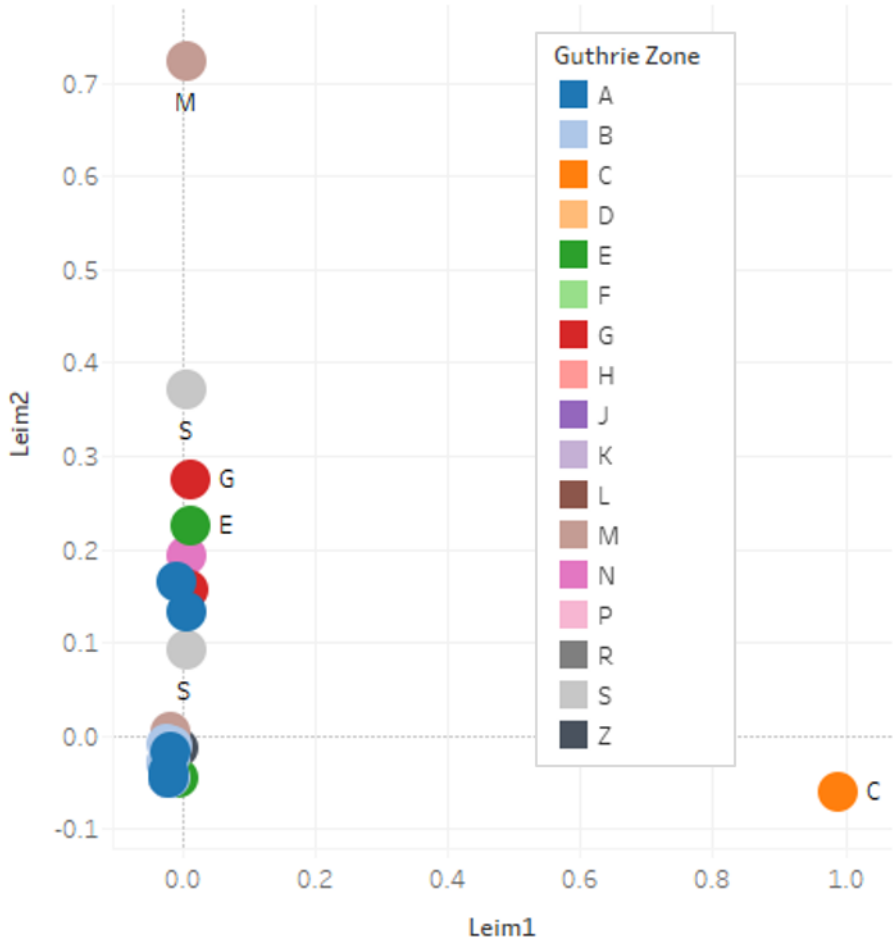

Figure 25: 2D LE scatter plot of outer joined data summarized by TaxaID.

The last two LE analyses yielded clearly spread placements of the datapoints. Figure 27 shows the results of the LE analysis using the inner joined data summarize by Guthrie zone. There is some agreement to the combined migratory model. Moving from right to left depicts the trajectory from northern zones such as A and B to the middle zones of K, M, and N after the "early split". However, the location of zones R and L are a slightly out of place in comparison to the combined migratory model.

The final LE analysis used the outer joined data summarized by Guthrie zone.

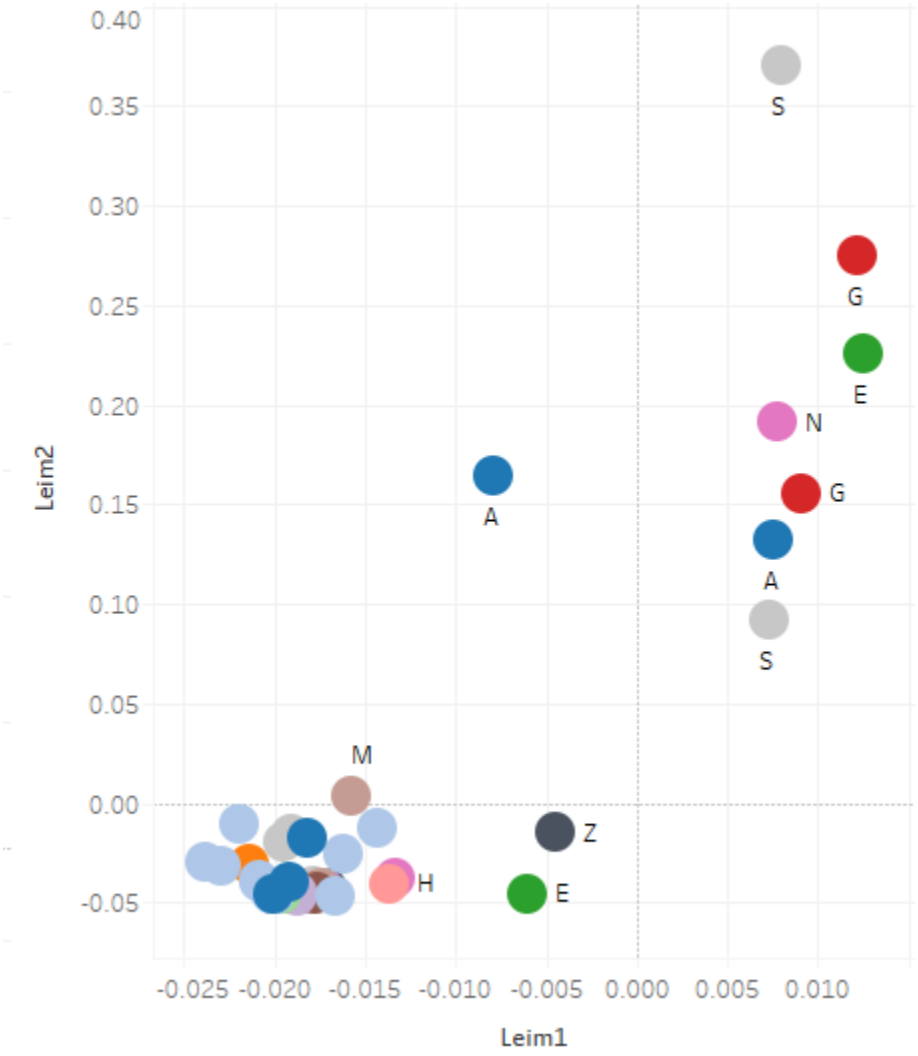

Figure 26: A zoomed-in view of figure 25, disregarding the farthest two points (C55Kele and M54Lamba).

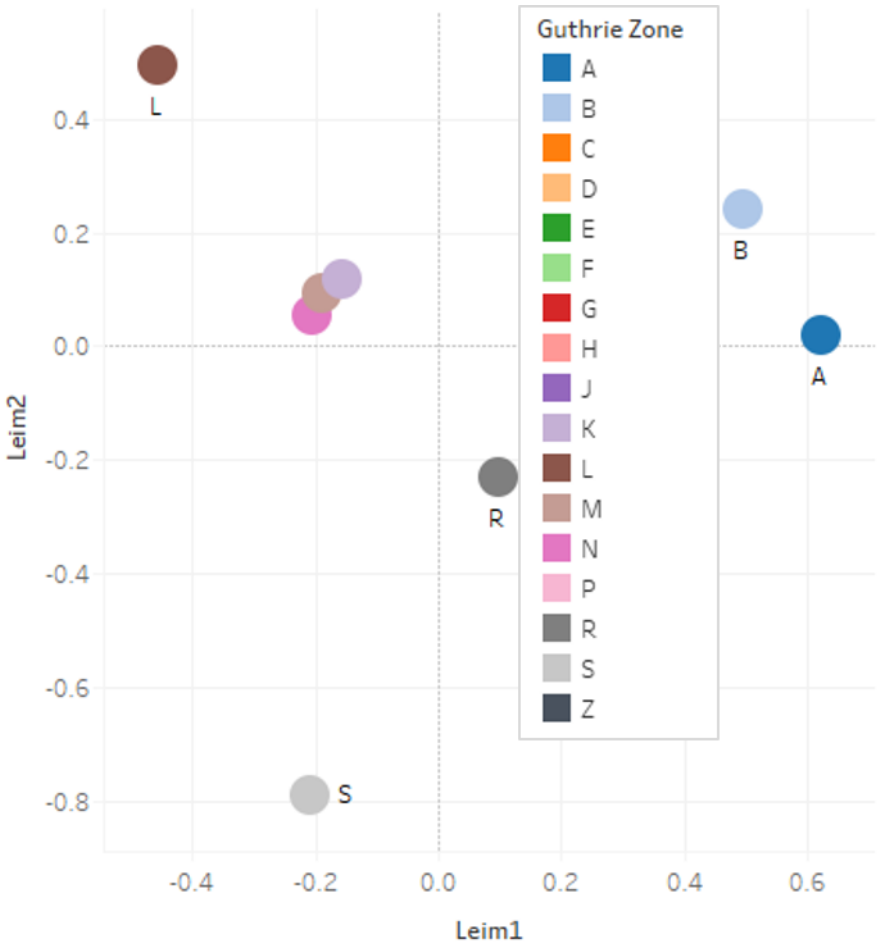

Figure 27: 2D LE scatter plot of inner joined data summarized by Guthrie zone.

This particular analysis resulted in the most noticeable agreement to the combined migratory model. See figure 28. In this visualization, there is an overall match of the trajectory. Starting at zones A and B and moving left, the next zones are C and H, which are the next two zones in the migratory model. Continuing left, zones E, F, and G are shown together, which matches the next area in the northwest corner of the map. Finally, ending with zones L, N, K, and R agrees with the final branches of the subsequent southeastern part of the migration seen in the model.

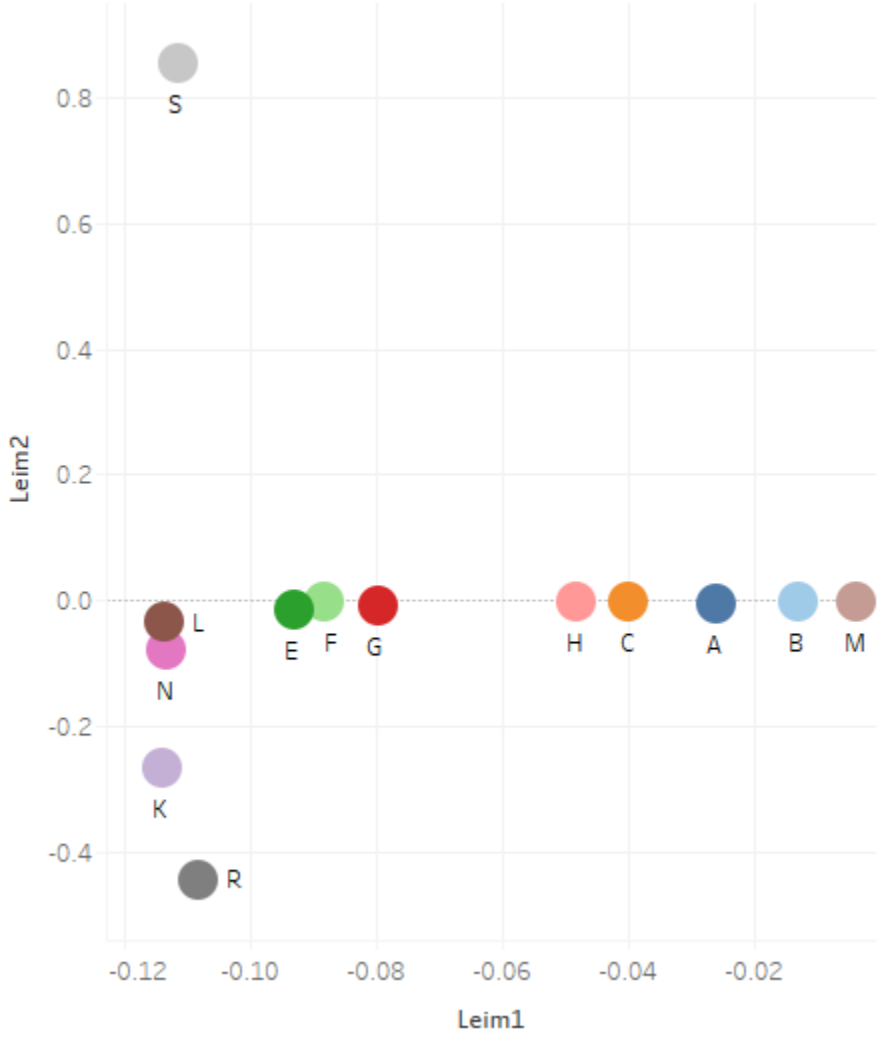

Figure 28: 2D LE scatter plot of outer joined data summarized by Guthrie zone.

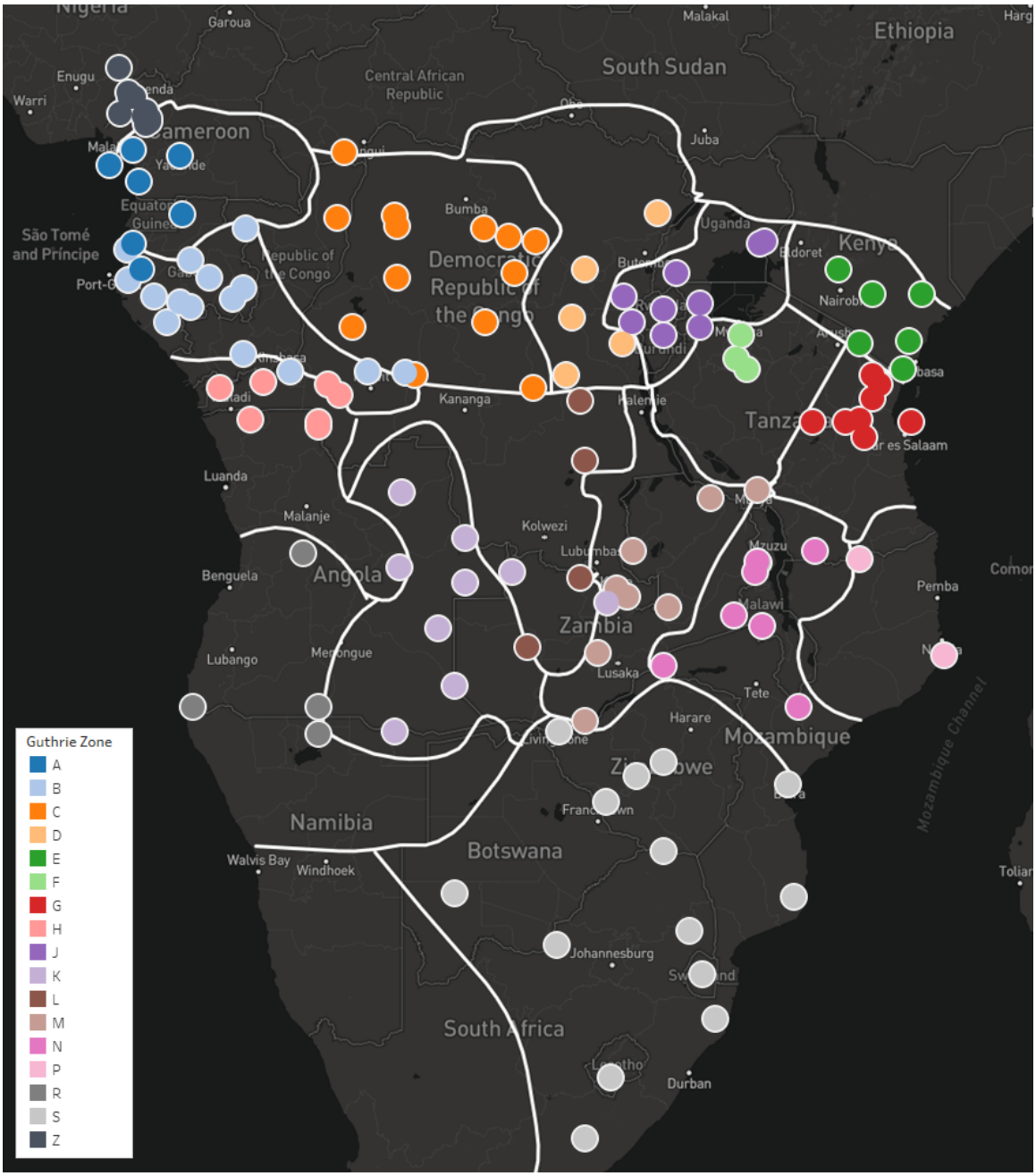

Figure 29: Map of TaxaID locations, colored by Guthrie zone.

Table 10: Stress values of the MDS Analyses. Note that all are below 0.05.

Join Type	**Summary Variable**	**Dimensions**	**Stress**
Inner	TaxaID	2	0.006988
Inner	TaxaID	3	0.008973
Inner	GuthrieZone	2	0.009335
Inner	GuthrieZone	3	0.005880
Outer	TaxaID	2	0.007755
Outer	TaxaID	3	0.009791
Outer	GuthrieZone	2	0.006923
Outer	GuthrieZone	3	0.009002

Table 11: Optimal k parameters in k-NN step for LE Analyses and their corresponding k-NN model accuracies.

Join Type	**Summary Variable**	**Dimensions**	**Optimal kNN**	**Accuracy**
Inner	TaxaID	2	14	0.6875
Inner	TaxaID	3	14	0.6875
Inner	GuthrieZone	2	2	0.7500
Inner	GuthrieZone	3	2	0.7500
Outer	TaxaID	2	8	0.6250
Outer	TaxaID	3	8	0.6250
Outer	GuthrieZone	2	12	0.7857
Outer	GuthrieZone	3	12	0.7857

3.4 Discussion

Using MDS and LE analyses proved to be both beneficial and insightful in understanding the validity of the combined Bantu migration model. Using two differently built datasets (the inner joined and outer joined data) as well as two ways of rolling up/summarizing the data (TaxaID and Guthrie zone) provided a very diverse set of output and resulting plots.

Many of these visualizations agree, at least in part, with the combined migratory model. It seems that, from the TaxaID-level data, there are TaxaIDs in the M Guthrie zone (such as M54Lamba) that seem to be outliers in relation to the other data points. For each plot, excluding or hiding these points enabled the other points to expand out and be viewed.

A possible reason for the unclear understanding of the inner joined-based plots is due to the low Guthrie zone coverage. For this dataset, only half of the Guthrie zones are available, which could negatively affect the resulting visualization.

In contrast, the outer joined dataset resulted in a much more robust set of datapoints on which to analyze and visualize. This is likely why the best agreement between any of the MDS or LE analyses (such as figure 28, for example) and the migratory model came from analyses using the larger, more encompassing dataset.

With this data quality notion in mind and given the evidence previously shown using the dimensionality reduction techniques, it appears that the combined migratory model using all three datasets has resulted in a highly confident model. This is backed by results of the dimensionality reduction analyses that are using the equivalent data

combing Ychr, mtDNA, and cultural aspects of the Bantu people.

CHAPTER 4: ASSESSMENT AND COMPARISON OF MODEL ACCURACY USING MACHINE LEARNING

4.1 Introduction

After generating the combined migratory model, its plausibility has been tested using dimensionality reduction from a visual standpoint. In addition, a more quantitative approach is necessary to compare and prove the combined model's power in relation to its constituent datasets' models.

In Chapter 2, a parsimonious tree search was completed, which generated both the individual models from one of the three datasets as well as the combined model. A competing method of tree generation that is often used is known as Maximum Likelihood. For maximum likelihood trees, the goal is to find the tree that maximizes the probability of n sequences ($x_{\cdot}$) given topology T and edge lengths $t_{\cdot}$ or $P(x_{\cdot}|T, t_{\cdot})$ [20].

This requires two steps to be completed many times and evaluated for the resultant likelihoods:

1. Search over the possible tree topologies, specifying the order of taxa assignments beforehand.

2. For each topology, search through all possible lengths of edges $t_{\cdot}$.

The main goal of a probability-based approach to constructing a phylogeny is to

evaluate trees according to their likelihood $P(data|tree)$ or, to take a more Bayesian view, look at their posterior probability $P(data|tree)$. Often, a secondary goal is also to find the likelihood of particular taxonomic features. However, these methods require a substitution model to serve as the basis for understanding the probability of changes in the sequences (mutations). A substitution model does not exist for data other than nucleic acids or amino acids. Thus, some other method must be used on the combined data given the inclusion of the Ychr STR and cultural information, which do not fit into the any available substitution models.

For Bantu, the specific goal is to understand the ancestral relationships and derive a migratory path that explains this lineage. Given a particular set of input data, is the data majorly indicative of the Guthrie zone to which the information belongs? Also, does the indication quality increase as more data is considered? If the combined model is a more comprehensive view of the Bantu people, their respective Guthrie zones, and thus their migratory path, a machine learning model can be trained such that the Guthrie zones of each record can be accurately predicted. The predictive power of the combined model should be stronger than that of the separate constituent datasets, assuming equivalent machine learning algorithms are used for training.

Using a random forest algorithm, machine learning models are generated to predict the Guthrie zone for each record of data. This is completed using each individual dataset as well as the combined dataset to compare the accuracy of the models. The overall accuracy of the combined model is statistically significantly higher than that of the other models.

Random decision forests were first created in 1995 by Tin Kam Ho from Bell

Laboratories [29]. The algorithm was then extended in 2001 by Leo Breiman and Adele Cutler to include a "bagging" method (also known as bootstrap aggregating) for random feature selection as well as a method for generating trees with a controlled variance [6]. The random forest algorithm is a supervised ensemble classification method that builds a multitude of decision trees. This algorithm is often preferred over singular decision trees as they correct against overfitting to the data. Decision trees (and thus decision forests) are non-parametric models, so they support data with varying distributions [50]. This is necessary for the diversity seen in this genetic and cultural sequence data.

4.2 Materials and Methods

Some of the data preprocessing steps that were performed for the dimensionality reduction analyses are reused for shaping the data in the correct format for machine learning. In figure 14, only steps 1 and 2 are necessary to shape the data for machine learning. This is due to the ability of a random forest algorithm to take in categorical data and missing data (as opposed to needing the numerical representation of the characters). For the R code to generate these datasets, see appendix D. For this analysis, five datasets are generated: the combined data, mtDNA data, Ychr data, Genetic data (containing both mtDNA and Ychr data), and Cultural data.

Each of these datasets are uploaded as separate saved datasets in the Microsoft Azure Machine Learning Studio experiment. From here, each saved dataset is brought onto the experiment canvas and connected to a Partition and Sample module. The Partition and Sample module is then connected, along with the Multiclass Decision

Forest model module, to the Cross Validate Model module. Finally, the Cross Validate Model module is connected to the Evaluate Model module, which will output the accuracy metrics for the trained and cross-validated model. See figure 30 for an example experiment workflow. Note that this is completed for each of the datasets in a single experiment. Each of the model outputs are downloaded and visualized for further comparison.

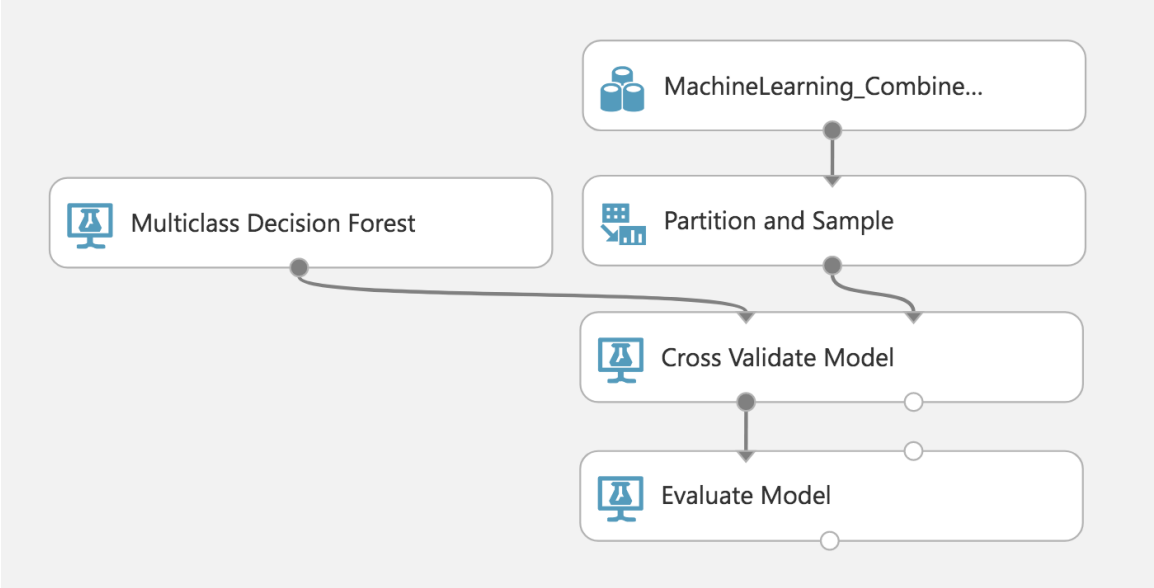

Figure 30: Example Azure Machine Learning experiment workflow.

4.2.1 Random Forest Model Generation

Using the Azure Machine Learning Studio, the individual and combined datasets are used to train individual Multiclass Decision Forest models. Using the input parameters listed in table 12, each of the five models are trained using the aforementioned corresponding datasets. The sets of trees are trained using the Guthrie zone as the dependent variable and then the individual characters of the sequence(s) as the features.

Table 12: Input parameters for the Azure Machine Learning Multiclass Decision Forest module.

Resampling Method	Bagging
Number of decision trees	250
Maximum depth of the decision trees	10000
Number of splits per node	250
Minimum number of samples per lead nodes	10

The algorithm randomly generates multiple decision trees, selecting different features each time. Then, the algorithm determines the classification output for each record by looking at the classification decision from each tree and picking the one with the greatest number of votes [50].

4.2.1.1 Cross-Validation

While cross-validation is not required for random forest models, it allows for visibility into the varying model performance with random subsets of data. For noisy data such as is seen in genetics, cross-validation will return performance metrics for each fold in the partitioned data to show an overall distribution of model accuracy, precision, recall, etc.

Using the Partition and Sample module, the data are separated into 3 folds for use in the subsequent cross-validation step. The parameter settings for the Partition and Sample module can been found in table 13.

Next, the fold indicator along with the rest of the data is passed to the Cross Validate Model module where the repetitive model generation occurs and is 3-fold cross-validated. The output nodes of this module include the prediction results as well as the cross-validation output metrics.

Finally, this cross-validated model is evaluated using the Evaluate Model module.

Table 13: Input parameters for the Azure Machine Learning Partition and Sample module

Partition or sample mode	Assign to Folds
Use replacement in the partitioning	True
Randomized split	True
Random seed	1337
Specify the partitioner method	Partition evenly
Specify the number of folds to split evenly into	3
Stratified split	True
Stratification key column	GuthrieZone

This module returns the accuracy metrics as well as the confusion matrix of the classification (Guthrie zone) prediction.

4.2.1.2 Statistical Test of Proportions

To statistically compare the results of the machine learning model accuracies, a test of equal proportions is performed. Using the *prop.test* function in base R, the accuracies of the individual models are compared to quantify if they are different enough to be statistically significant. The *prop.test* function performs an n-sample test of proportions and returns the χ^2 value and p-value as a result. This performs a z-test to test the equivalency of the input proportions. That is, the null tested is that the proportions in each group are the same. In this case, each proportion is the relationship of correct predictions versus the total number of records in each machine learning model.

4.3 Results

The overall accuracy of each of the random forest models corresponding to an individual input data can be seen in table 14. The machine learning model that uses the combined dataset has the highest overall accuracy. The model using the Ychr

Table 14: Overall random forest 3-fold cross-validation accuracy by model.

Model	Overall Accuracy
Combined	72.51%
mtDNA	42.84%
Ychr	69.46%
Genetic (mtDNA + Ychr)	55.22%
Cultural	52.02%

dataset also has a similarly high accuracy, with the remaining models having much lower overall accuracies.

To visually compare the classification accuracy by Guthrie zone, a confusion matrix was generated. In figure 31, the combined model shows the highest accuracy (correct classifications) along the diagonal of the matrix. However, the Ychr and Cultural models also have strong accuracies. It should be noted that it seems the many records in all the models are often misclassified as the S Guthrie zone. Also note that for some zones, there may not be any accurate predictions. For example, Guthrie zone A was never predicted to be zone A in the mtDNA model. This influences the following metrics' visualizations.

Note that there are certain zones with no correct predictions in the data. These are zones with low coverage in the training dataset, leading to less predictive accuracy in zones E, F, G, and N. See table 15.

The distribution of model accuracy can be examined at a lower level by showing the distribution by either Guthrie zone or the cross-validation fold number.

In figure 32, the top two models are once again the combined model and the Ychr model. Both of these models have a higher accuracy in more of the Guthrie zones. If the zones with low data coverage are are excluded, the combined model becomes

Confusion Matrix

Model	Guthrie Zone	Scored Labels A	B	C	F	H	K	L	M	N	R	S
Combined	A	43.37%	14.46%				16.27%					25.90%
	B		35.66%			20.52%	15.14%					28.69%
	C			40.83%		7.82%	20.05%					31.30%
	E											100.00%
	F				78.26%							21.74%
	G						50.00%				50.00%	
	H					61.78%	8.07%					30.15%
	K		0.17%				70.14%					29.69%
	L						70.09%	0.43%	1.28%			28.21%
	M						18.88%		51.45%			29.67%
	N						15.38%				46.15%	38.46%
	R						3.90%				66.92%	29.18%
	S						2.02%				0.16%	97.82%
Cultural	A	68.75%										31.25%
	B		66.67%									33.33%
	E	50.00%									50.00%	
	F				69.57%							30.43%
	G	100.00%										
	H					75.00%						25.00%
	K					34.69%	38.78%					26.53%
	L						34.78%	26.09%				39.13%
	M							33.35%	36.46%			30.20%
	N				50.00%					12.50%	12.50%	25.00%
	R									33.35%	36.99%	29.66%
	S										33.37%	66.63%
Genetics	A					20.45%					31.82%	47.73%
	B		51.63%			17.93%					0.82%	29.62%
	C		4.65%	64.30%		2.20%						28.85%
	H		2.00%			65.61%						32.39%
	K						19.80%				5.04%	75.15%
	L						1.57%				5.70%	92.74%
	M						0.99%		44.69%		6.24%	48.09%
	N										8.33%	91.67%
	R			0.62%			1.10%				39.76%	58.52%
	S					0.02%	4.12%				3.83%	92.02%
mtDNA	A			29.85%		21.64%					28.36%	20.15%
	B		45.92%	30.16%		23.91%						
	C		3.42%	73.59%		3.42%						19.56%
	E											100.00%
	G											100.00%
	H		2.67%	29.55%		64.61%						3.17%
	K			29.81%		0.20%	17.30%				5.01%	47.68%
	L			30.77%							1.14%	68.09%
	M			29.31%			1.04%		43.66%		6.52%	19.45%
	N			23.08%							7.69%	69.23%
	R			30.86%			0.75%				35.34%	33.06%
	S			29.69%		0.05%	4.05%				4.60%	61.61%
Ychr	A	28.05%	19.51%			1.83%			17.07%		32.32%	1.22%
	B	4.58%	30.68%			3.59%	2.99%		25.50%		32.07%	0.60%
	C	0.73%	3.67%	65.28%		0.49%			1.22%		28.61%	
	F										100.00%	
	H	1.15%				68.86%	0.33%				29.32%	0.33%
	K						55.54%	0.29%	0.81%		37.22%	6.13%
	L						5.13%	46.01%	1.85%		37.32%	9.69%
	M						2.35%	0.43%	62.08%		33.29%	1.86%
	N						33.33%		8.33%		50.00%	8.33%
	R						1.72%				96.02%	2.27%
	S						2.25%	0.08%	0.01%		31.44%	66.22%

Figure 31: Confusion matrix of Guthrie zone predictions versus actual zones by model.

Table 15: Zone coverage in the Random Forest model training datasets.

Guthrie Zone	**Combined**	**Cultural**	**mtDNA**	**Ychr**	**Genetic**	***Overall***
A	0.67%	0.57%	0.55%	0.67%	0.54%	*0.60%*
B	2.03%	0.48%	1.51%	2.04%	1.51%	*1.58%*
C	1.66%	0.00%	1.67%	1.66%	1.68%	*1.42%*
E	0.01%	0.01%	0.01%	0.00%	0.00%	*0.01%*
F	0.09%	0.14%	0.00%	0.09%	0.00%	*0.06%*
G	0.01%	0.01%	0.01%	0.00%	0.00%	*0.00%*
H	2.46%	0.05%	2.45%	2.47%	2.46%	*2.10%*
K	16.62%	0.29%	16.76%	16.55%	16.70%	*14.24%*
L	2.84%	0.14%	2.87%	2.85%	2.88%	*2.46%*
M	13.97%	14.99%	14.12%	14.01%	14.16%	*14.20%*
N	0.05%	0.05%	0.05%	0.05%	0.05%	*0.05%*
R	24.30%	33.50%	24.58%	24.37%	24.65%	*25.80%*
S	35.29%	49.79%	35.41%	35.24%	35.37%	*37.46%*
Grand Total	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*	100.00%

evident that it is more accurate (in the zones that had an acceptable number of training instances). See figure 33.

Furthermore, the additional metrics of average log loss, precision, and recall are visualized in box plot format as well. See figure 34. The lowest average log loss belongs to the genetic model, with the Ychr and combined models following. The highest precision is given by the Ychr model.

The accuracy distributions are also visualized by the cross-validation folds. Given that only 3-fold cross-validation was completed, each model only has three points from which to plot. As seen in figure 35, the combined and Ychr models have similar accuracies over the three folds with the Ychr model marginally having the highest accuracy fold between the two. The combined model, though, has a lower range of accuracy, therefore showing that the combined model has the superior accuracy overall.

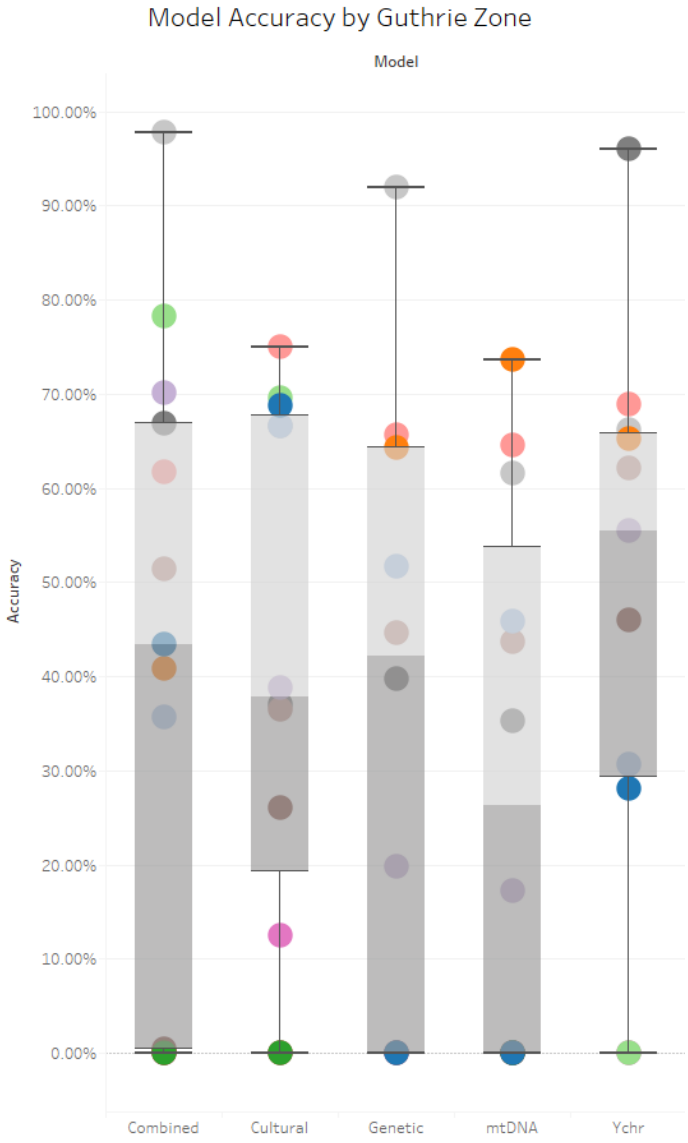

Figure 32: Box plots of accuracy by model, colored by Guthrie zone.

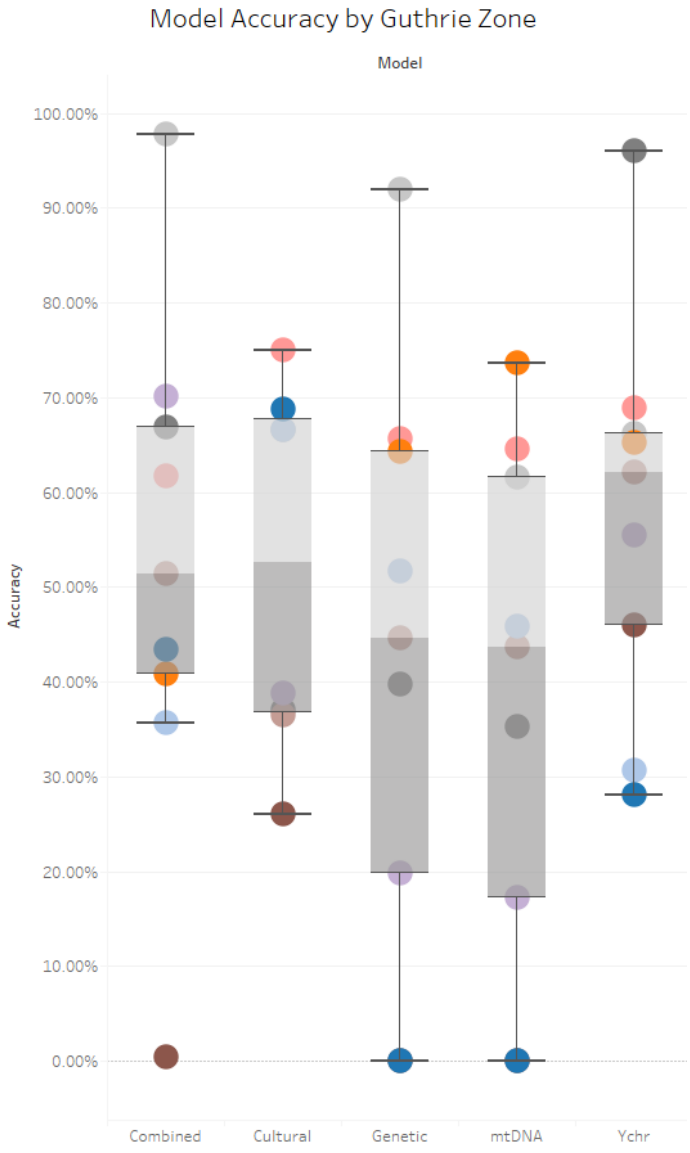

Figure 33: Box plots of accuracy by model, colored by Guthrie zone. Shown excluding zones E, F, G, and N, which had low observation counts in the training datasets for these models.

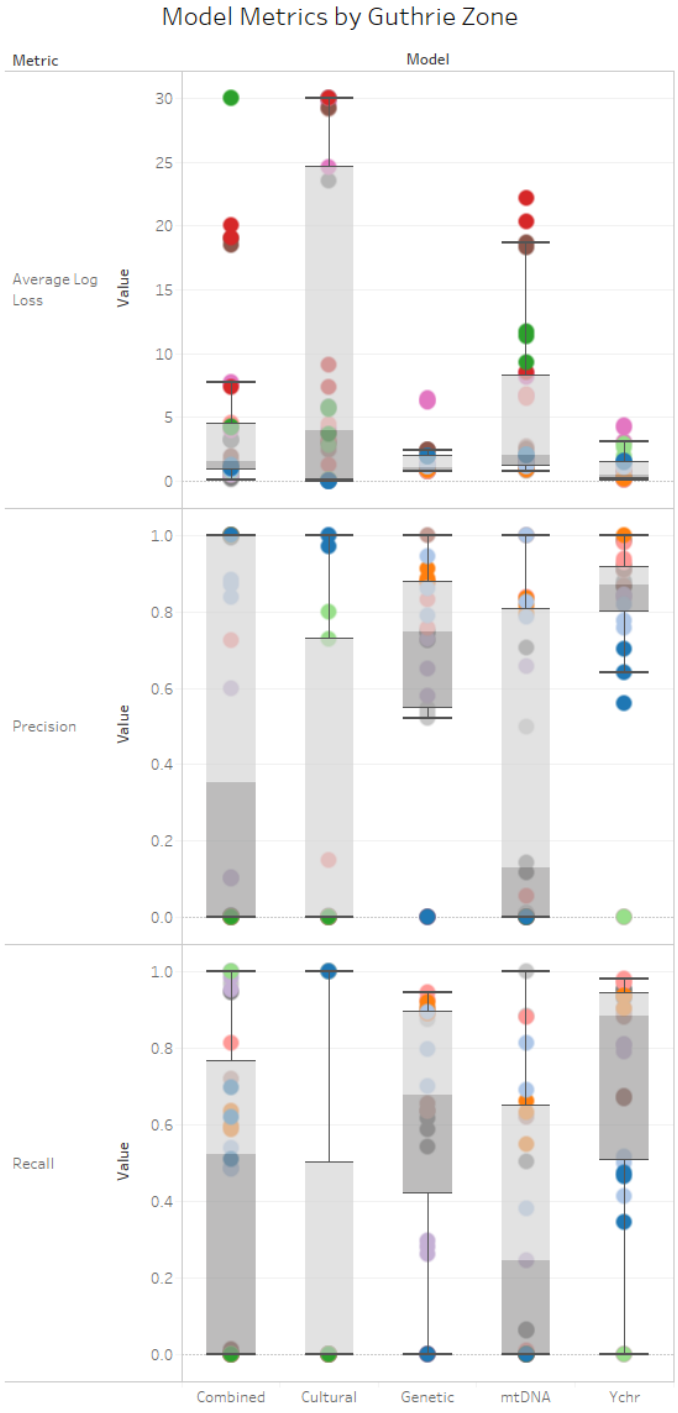

Figure 34: Box plots of additional model metrics, colored by Guthrie zone.

Table 16: R output of the 5-proportion Z-test for equivalence of model accuracies.

5-sample test for equality of proportions without continuity correction	
prop.test(x = ModelAccuracy$Correct, n = ModelAccuracy$Total)	
data: ModelAccuracy$Correct out of ModelAccuracy$Total	
X-squared:	6124.5
df:	4
p-value:	<2.2e-16
alternative hypothesis:	two.sided
sample estimates:	
prop 1	0.6945880
prop 2	0.4283911
prop 3	0.5202384
prop 4	0.7251017
prop 5	0.5521931

For average log loss, precision, and recall, the superior model is less clear. In figure 36, the genetic, Ychr, and combined models have the lowest distributions of average log loss. For precision and recall, the outcome is similar to before. It appears that the Ychr has the highest precision and recall distributions. This is due to the overall distributions of these metrics being skewed by TaxaIDs with 0% precision or recall. However, for TaxaIDs with precision/recall values $\neq 0$, the combined, Ychr, and genetic models all show acceptable levels. The mtDNA model has the worst distributions of these metrics overall.

The 5-sample proportion test, shown in table 16, rejects the null hypothesis that all the proportions are equal. In other words, the overall accuracy of the individual models are statistically significantly different from one another. More importantly, the combined model's highest overall accuracy is statistically different than that of the Ychr model.

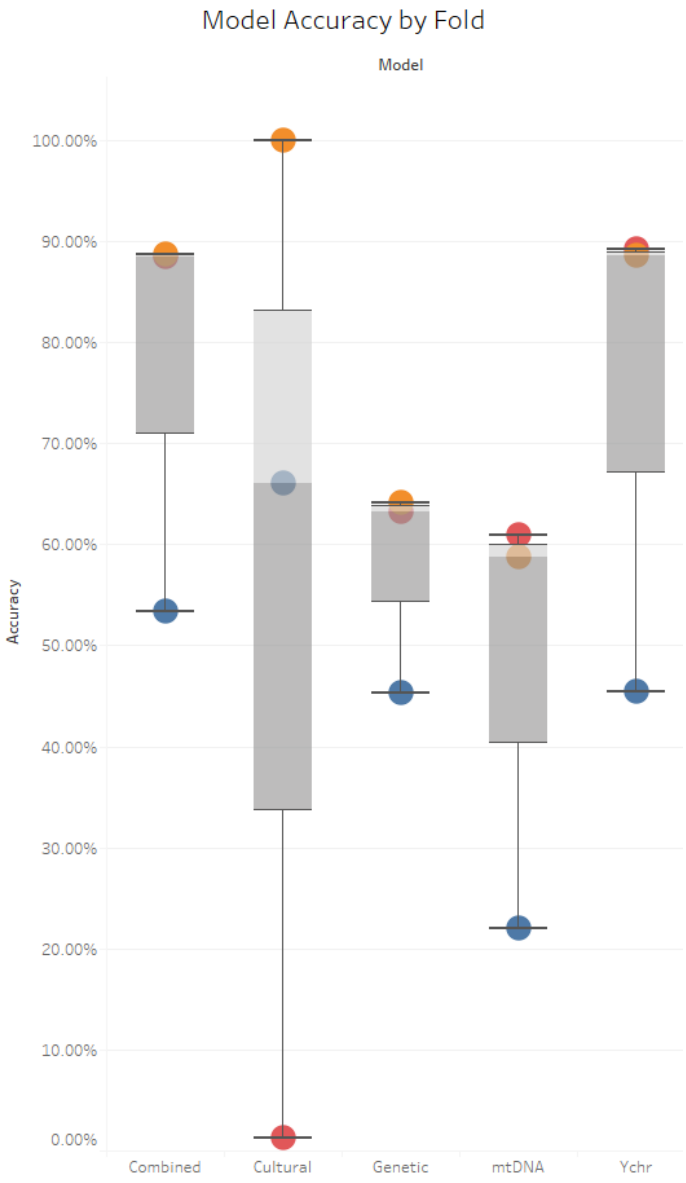

Figure 35: Box plots of accuracy by model, colored by cross-validation fold.

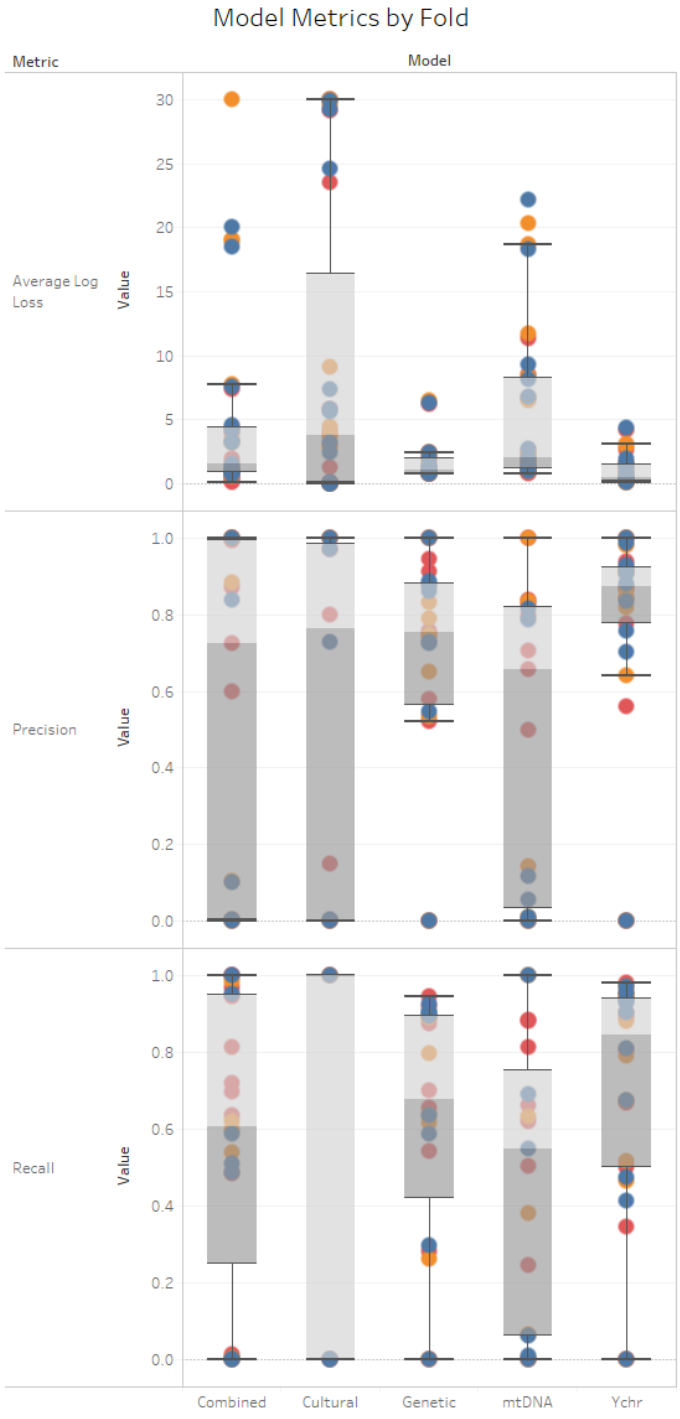

Figure 36: Box plots of additional model metrics, colored by cross-validation fold.

4.4 Discussion

Using the random forest algorithm to generate predictive models for each of the datasets provided new insight into the predictive power of each type of data. Comparing the overall accuracies, the distribution of accuracies, and other metrics supports the importance of using the combined data. Given the superior overall accuracy and distribution of accuracies by both Guthrie zone and fold, the combined model is proven to be more indicative of the Guthrie zone. The statistical test also confirms that the difference in overall accuracies is statistically significant.

While there are some metrics that show competing models may have a better precision, recall, or average log loss, this is likely due to the non-classification of certain Guthrie zones in the data. Without any records being predicted to belong in a particular zone, the precision and recall are shown to be 0%, skewing the overall distribution. Excluding these for all the models results in a clearer box plot comparison, one where the combined or Ychr models most often have the best distribution of the metric in question.

This confirms that the migratory model using the combined data has a higher indicative power to explain more of the variation in the migration trajectory from zone to zone. Using these machine learning random forest models, this builds confidence that the migratory model generated by the combined data is a more confident overall model than other models using a singular dataset.

CHAPTER 5: CONCLUSION

Understanding the evolutionary relationships of a set of organisms is often a difficult problem as getting enough data to derive a clear tree is tricky. Furthermore, extending this work geospatially to determine a migratory path is often an arduous task in and of itself. For humans, the task is even less straightforward because of the different aspects we pass along to our offspring. We not only pass down our genetic information, but the cultural norms we practice. 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 taught how to behave by them as well.

Beyond the first step creating a combined phylogeny, the data here have been further analyzed using dimensionality reduction and machine learning methods. Looking at the results from the different analyses helped to build confidence around the resultant migratory model. This suggests that the combined model most confidently follows the true Bantu migration path (given the available data) more closely than the individual models (models generated using a singular dataset) or previous methods' models (models postulated by other researchers). This led to the partial support for the initial "early split" hypothesis, but poses evidence of a subsequent migration.

Visualization also played an integral part in the interpretation of the analysis results. Plotting the phylogenetic tree and the apomorphies geographically proved to

be imperative in the derivation of the migratory model. Plotting the phylogeny of the combined dataset was necessary in determining the areas of agreement or disagreement to the other, previously-published models. Plus, representing Guthrie zone transitions as a geographic migration trajectory was the final step in generating the assessment of the migratory path. In addition, the results of the dimensionality reduction and machine learning exercises were graphically represented for easier interpretation and comparison to the phylogenetic analyses.

All steps of this research were designed to test the hypothesis that the Bantu expansion can be characterized by a primary split in lineages. Given the results of the overall analysis, the "early split" hypothesis is supported, but the results also provide evidence for a subsequent migration that is not shown in the published migratory models previously referenced. This alludes to the notion that the migration is more complicated than originally presumed. The de Filippo and Grollemund models show a rather linear path in which the individuals moved in a mostly singular direction from the northwest to the southeast. However, the results of this work show a more complex movement characterized by a west to east trajectory subsequently followed by a north to south branching expansion. This is most similar to the Currie model superficially, but is still a very different trajectory overall. See figure 1.

The completion of this research has shown the versatility and usefulness of integrated data, phylogenetics, and visual analytics in bioinformatics and computational biology, specifically in the examination of the Bantu migration. This has provided a more complete look at Bantu migration, encompassing both cultural and genetic viewpoints.

5.1 New Working Hypothesis

Initially, the hypothesis being tested was that the Bantu expansion can be characterized by a primary split in lineages. Given the results of the analyses performed in this work, the initial hypothesis cannot be fully rejected, but it seems that the hypothesis is incomplete. Also, it seems that the split did not occur quite as "early" as originally hypothesized.

A better working hypothesis would be that the migration began in present day southern Cameroon and experienced a binary split early on, but somewhere more eastward such as in the northern part of present day Democratic Republic of the Congo (Guthrie zone D). Then, subsequent migrations occurred in a more branching pattern south throughout the rest of the continent.

This new working hypothesis can be supported by the results of the previously highlighted analyses.

5.1.1 Evidence of Early Split

The result of the parsimonious phylogenetic tree first shows zone transitions between zones A, B, and H. This looks most similar to the first portion of the de Filippo model (b) in figure 1. Then, the transitions move north and then east from zone H to C to D. This characterizes the linear portion of the migration.

At zone D, the split occurs and there are then branches going from D eastward (to zones J, F, G, and E) and then a southward trajectory to zone L and branching out to the rest of the more southern zones next.

The sequence of these transitions provide evidence partially confirming the "early

split" hypothesis.

5.1.2 Evidence of Subsequent Migration

Following the expansion to zone L, the migratory path branches out significantly from there. This is the portion of the model that was not captured by the original hypothesis in that this is a subsequent migration following the initial split.

From L, there are three overall directions of expansion that can be identified: to the southwest, to the east, and to the south. The southwest movement is characterized by transitions from zone L to K to R. The southward movement is characterized by transitions from zone L to M to S. These are paths that would likely not be as easily identified by previous models because of the theory that movement occurs mainly east to west due to the avoidance of encountering differing climates and habitats. However, there are quite a few cases in this model that do not follow "path of least resistance" theory. See figure 37. Finally, the subsequent eastward movement is characterized by transitions from zone L to M to N to P.

The radiating trajectories after the initial "early split" provide evidence in a more complex migration than originally assumed. These branching events are not shown in any of the previously published models. Furthermore, the other models show far less activity the southern zones than the model generated here. Thus, there is further work to understand the later migratory events shown here in the southernmost transitions.

5.2 Significance and Future Work

The Bantu migration is important in the understanding of human history and evolution as it marks one of the most influential cultural events of all time. This

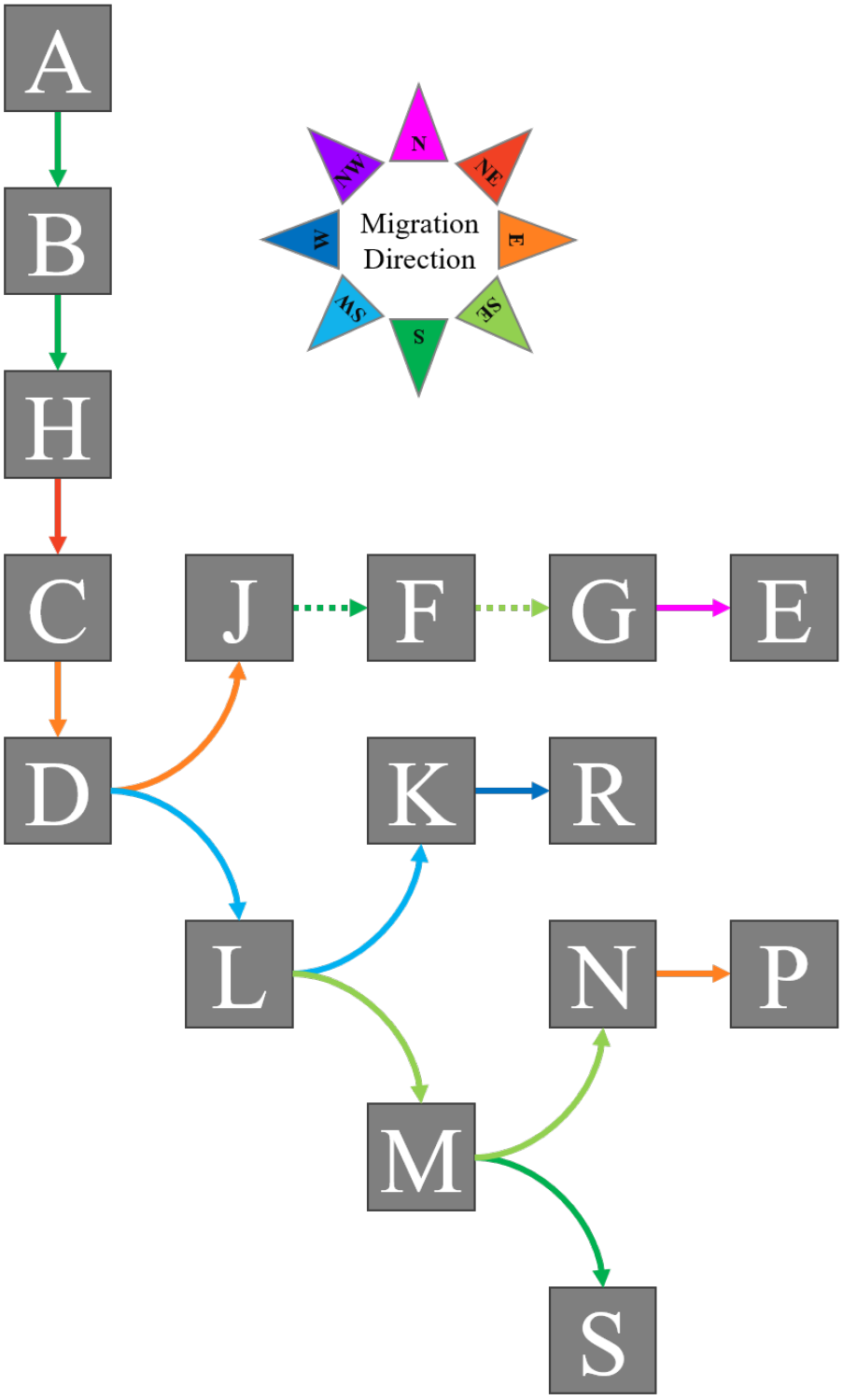

Figure 37: Color-coded migration directions between Guthrie zones.

work is significant in that it provides another perspective of the migration using both genetic and cultural data. Contrary to previously published models, it seems that the migration is much more complicated than originally thought. Future work is at hand to further understand the timeline around this migration and how other events affected it. Also, more understanding is needed around the effect of geographical features on the migration path. Specifically, geographical features such as swamps, forests, and mountains seem to have played a role in the "early split" hypothesis and the subsequent migration in that individuals settled around these geographical features rather than in them.

5.2.1 Applicability to Other Research

Beyond Bantu migration, the pipeline designed in this work can be applied to many other areas of research. For any organismal migration, this same analysis can be accomplished given access to the appropriate data. In infectious disease research in particular, analyzing diverse data to get a full picture of the movement of organisms is a crucial step. For example, mapping the infection path of serious pathogens such as Middle East respiratory syndrome [30], Zika virus [15], or influenza virus [31] has proven effective from an epidemiological standpoint.

This pipeline also helps to address the complicated task of analyzing data when the data is both diverse and difficult to combine. In any research setting where data is diverse, the first challenge is to combine the data in a meaningful way and the second is to analyze it holistically without compromising parts of the data due to limitations in software, hardware, or understanding. In this research, it has been demonstrated

that large amounts of diverse data can be successfully integrated and analyzed, both phylogenetically and beyond. Then, the output can be visualized for a geographic view of the analysis. These steps prove to be important in the overall geospatial understanding of organismal movement.

REFERENCES

[1] C. Barbieri, M. Vicente, S. Oliveira, K. Bostoen, J. Rocha, M. Stoneking, and B. Pakendorf. Migration and interaction in a contact zone: mtdna variation among bantu-speakers in southern africa. *PLOS ONE*, 9(6):1–14, 06 2014.

[2] D. M. Behar, R. Villems, H. Soodyall, J. Blue-Smith, L. Pereira, E. Metspalu, R. Scozzari, H. Makkan, S. Tzur, D. Comas, and et al. The dawn of human matrilineal diversity. *The American Journal of Human Genetics*, 82(5):11301140, 2008.

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

[4] G. Berniell-Lee, F. Calafell, E. Bosch, E. Heyer, L. Sica, P. Mouguiama-Daouda, L. V. D. Veen, J.-M. Hombert, L. Quintana-Murci, D. Comas, and et al. Genetic and demographic implications of the bantu expansion: Insights from human paternal lineages. *Molecular Biology and Evolution*, 26(7):15811589, 2009.

[5] I. Borg and P. J. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[6] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[7] N. Brucato, O. Cassar, L. Tonasso, P. Tortevoye, F. Migot-Nabias, S. Plancoulaine, E. Guitard, G. Larrouy, A. Gessain, J.-M. Dugoujon, and 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):314, 2010.

[8] G. v. Brummelen. *Heavenly mathematics: the forgotten art of spherical trigonometry*. Princeton University Press, 2013.

[9] J. Butler. Addressing y-chromosome short tandem repeat (y-str) allele nomenclature. 4:125–148, 01 2008.

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

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

[12] L. Castr, S. Tofanelli, P. Garagnani, C. Bini, X. Fosella, S. Pelotti, G. Paoli, D. Pettener, and D. Luiselli. mtdna variability in two bantu-speaking populations (shona and hutu) from eastern africa: Implications for peopling and migration patterns in sub-saharan africa. *American Journal of Physical Anthropology*, 140(2):302311, 2009.

[13] M. Coelho, F. Sequeira, D. Luiselli, S. Beleza, and J. Rocha. On the edge of bantu expansions: mtdna, y chromosome and lactase persistence genetic variation in southwestern angola. *BMC Evolutionary Biology*, 9(1):80, Apr 2009.

[14] 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.

[15] A. de Bernardi Schneider, R. W. Malone, J.-T. Guo, J. Homan, G. Linchangco, Z. L. Witter, D. Vinesett, L. Damodaran, and D. A. Janies. Molecular evolution of zika virus as it crossed the pacific to the americas. *Cladistics*, 33(1):1–20, 2017.

[16] 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.

[17] C. de Filippo, P. Heyn, L. Barham, M. Stoneking, and B. Pakendorf. Genetic perspectives on forager-farmer interaction in the luangwa valley of zambia. *American Journal of Physical Anthropology*, 141(3):382–394, 2010.

[18] J. Diamond. Guns, germs, and steel: The fates of human societies norton, ww & company. *Inc. Sales*, 1997.

[19] J. Diamond and P. Bellwood. Farmers and their languages: The first expansions. *Science*, 300(5619):597–603, 2003.

[20] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 2013.

[21] J. Fujihara, I. Yuasa, T. Muro, R. Iida, E. Tsubota, H. Nakamura, S. Imamura, T. Yasuda, and H. Takeshita. Allele frequencies and haplotypes for 28 y-strs in ovambo population. *Legal Medicine*, 11(4):205208, 2009.

[22] M. K. Gonder, H. M. Mortensen, F. A. Reed, A. de Sousa, and S. A. Tishkoff. Whole-mtdna genome sequence analysis of ancient african lineages. *Molecular biology and evolution*, 24(3):757–768, 2006.

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

[24] 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.

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

[26] M. Guthrie. *Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages*, volume 1-4. Gregg International, 1967-1971.

[27] B. M. Henn, C. Gignoux, A. A. Lin, P. J. Oefner, P. Shen, R. Scozzari, F. Cruciani, S. A. Tishkoff, J. L. Mountain, and P. A. Underhill. Y-chromosomal evidence of a pastoralist migration through tanzania to southern africa. *Proceedings of the National Academy of Sciences*, 105(31):10693–10698, 2008.

[28] L. Henry and H. Wickham. *purrr: Functional Programming Tools*, 2017. R package version 0.2.3.

[29] T. K. Ho. Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, page 278282, Aug 1995.

[30] D. Janies, C. T. Ford, and L. Damodaran. Spread of middle east respiratory coronavirus: Genetic versus epidemiological data. *Online Journal of Public Health Informatics*, 9(1), Feb 2017.

[31] D. A. Janies, L. W. Pomeroy, C. Krueger, Y. Zhang, I. F. Senturk, K. Kaya, and . V. atalyrek. Phylogenetic visualization of the spread of h7 influenza a viruses. *Cladistics*, 31(6):679–691, 2015.

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

[33] G. A. Korn and T. M. Korn. *Mathematical handbook for scientists and engineers: definitions, theorems, and formulas for reference and review*. Dover, 2000.

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

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

[36] J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, Jun 1964.

[37] N. Leat, L. Ehrenreich, M. Benjeddou, K. Cloete, and S. Davison. Properties of novel and widely studied y-str loci in three south african populations. *Forensic Science International*, 168(2):154 – 161, 2007.

[38] M. Lecerf, M. Filali, G. Grsenguet, A. Ndjoyi-Mbiguino, J. L. Goff, P. D. Mazancourt, and L. Blec. Allele frequencies and haplotypes of eight y-short tandem repeats in bantu population living in central africa. *Forensic Science International*, 171(2-3):212215, 2007.

[39] 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.

[40] P. Mellars. Going east: New genetic and archaeological perspectives on the modern human colonization of eurasia. *Science*, 313(5788):796–800, 2006.

[41] D. Mishmar, E. Ruiz-Pesini, P. Golik, V. Macaulay, A. G. Clark, S. Hosseini, M. Brandon, K. Easley, E. Chen, M. D. Brown, R. I. Sukernik, A. Olckers, and D. C. Wallace. Natural selection shaped regional mtdna variation in humans. *Proceedings of the National Academy of Sciences*, 100(1):171–176, 2003.

[42] K. Neumann, K. Bostoen, A. Hhn, S. Kahlheber, A. Ngomanda, and B. Tchiengu. First farmers in the central african rainforest: A view from southern cameroon. *Quaternary International*, 249:53 – 62, 2012. Long-term perspectives on human occupation of tropical rainforests.

[43] R. Oliver. The problem of the bantu expansion. *The Journal of African History*, 7(3):361–376, 1966.

[44] L. Quintana-Murci, H. Quach, C. Harmant, F. Luca, B. Massonnet, E. Patin, L. Sica, P. Mouguiama-Daouda, D. Comas, S. Tzur, O. Balanovsky, K. K. Kidd, J. R. Kidd, L. van der Veen, J.-M. Hombert, A. Gessain, P. Verdu, A. Froment, S. Bahuchet, E. Heyer, J. Dausset, A. Salas, and D. M. Behar. Maternal traces of deep common ancestry and asymmetric gene flow between pygmy hunter-gatherers and bantu-speaking farmers. *Proceedings of the National Academy of Sciences of the United States of America*, 105 5:1596–601, 2008.

[45] W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2017. R package version 1.7.5.

[46] 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.

[47] D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.

[48] T. Software. Tableau. 2017.

[49] D. B. Suits. Use of dummy variables in regression equations. *Journal of the American Statistical Association*, 52(280):548–551, 1957.

[50] J. Takaki, T. Petersen, and G. Ericson. Multiclass decision forest - azure machine learning studio, Jan 2018.

[51] S. A. Tishkoff, M. K. Gonder, B. M. Henn, H. Mortensen, A. Knight, C. Gignoux, N. Fernandopulle, G. Lema, T. B. Nyambo, U. Ramakrishnan, F. A. Reed, and J. L. Mountain. History of click-speaking populations of africa inferred from mtdna and y chromosome genetic variation. *Molecular Biology and Evolution*, 24(10):2180–2195, 2007.

[52] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[53] 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.

[54] 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.

[55] H. Wickham, R. Francois, L. Henry, and K. Mller. *dplyr: A Grammar of Data Manipulation*, 2017. R package version 0.7.2.

[56] H. Wickham and L. Henry. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2017. R package version 0.7.0.

[57] H. Wickham, J. Hester, and R. Francois. *readr: Read Rectangular Text Data*, 2017. R package version 1.1.1.

[58] E. O. Wiley. *Phylogenetics. The theory and practice of phylogenetic systematics*. Wiley XV, 1981.

APPENDIX A: DATABASE SETUP CODE

A Create Table SQL Scripts

Listing 1: Create mtDNA table

```
CREATE TABLE [dbo].[mtDNA](
        [Taxa] [nvarchar](255) NULL,
        [GuthrieZone] [nvarchar](10) NULL,
        [BCMID] [nvarchar](255) NULL,
        [TaxaID] [nvarchar](255) NULL,
        [SampleID] [nvarchar](255) NULL,
        [mtDNA] [nvarchar](max) NULL
)
```

Listing 2: Create Ychr table

```
CREATE TABLE [dbo].[Ychr](
        [Taxa] [nvarchar](255) NULL,
        [GuthrieZone] [nvarchar](10) NULL,
        [BCMID] [nvarchar](255) NULL,
        [TaxaID] [nvarchar](255) NULL,
        [SampleID] [nvarchar](255) NULL,
        [Ychr_STR] [nvarchar](255) NULL
)
```

Listing 3: Create Cultural table

```
CREATE TABLE [dbo].[Cultural](
        [Taxa] [nvarchar](255) NULL,
        [TaxaID] [nvarchar](255) NULL,
        [BCMID] [nvarchar](255) NULL,
        [GuthrieZone] [nvarchar](10) NULL,
        [Cultural_EthnogAtlas] [nvarchar](255) NULL
)
```

Listing 4: Create Geographic Information table

```
CREATE TABLE [dbo].[GeographicInformation](
        [GuthrieZone] [nvarchar](255) NULL,
        [BCM ID_FULL] [nvarchar](255) NULL,
        [BCM ID_SIMPLE] [nvarchar](255) NULL,
        [BCM ID_SUFFIX] [nvarchar](255) NULL,
        [BCM ID_INDEX] [int] NULL,
        [BCM Name] [nvarchar](255) NULL,
        [BCM FULL NAME] [nvarchar](255) NULL,
        [Long_Orig] [nvarchar](255) NULL,
        [Lat_Orig] [nvarchar](255) NULL,
        [Latitude] [float] NULL,
        [Longitude] [float] NULL
)
```

B Create View SQL Scripts

Listing 5: Create Inner Joined Dataset View

```
CREATE VIEW [dbo].[vwInnerJoinedData] AS(
SELECT          c.TaxaID
                ,c.GuthrieZone
                ,c.Taxa AS CulturalTaxa
                ,mt.Taxa AS mtDNATaxa
                ,y.Taxa AS YchrTaxa
                ,c.Cultural_EthnogAtlas
                ,mt.mtDNA
                ,y.Ychr_STR
FROM Cultural c
JOIN mtDNA mt
ON mt.TaxaID = c.TaxaID
JOIN Ychr y
ON y.TaxaID = c.TaxaID
AND mt.TaxaID = y.TaxaID)
```

Listing 6: Create Outer Joined Dataset View

```
CREATE VIEW [dbo].[vwOuterJoinedData] AS(
SELECT  (CASE WHEN mt.TaxaID IS NULL THEN
            (CASE WHEN y.TaxaID IS NULL THEN
```

```
        c.TaxaID ELSE y.TaxaID END) ELSE
        mt.TaxaID END) AS TaxaID
            ,(CASE WHEN mt.GuthrieZone IS NULL THEN
                    (CASE WHEN y.GuthrieZone IS NULL THEN
                    c.GuthrieZone ELSE y.GuthrieZone END)
                    ELSE mt.GuthrieZone END) AS GuthrieZone
            ,mt.Taxa AS mtDNATaxa
            ,y.Taxa AS YchrTaxa
            ,c.Taxa AS CulturalTaxa
            ,mt.mtDNA
            ,y.Ychr_STR
            ,c.Cultural_EthnogAtlas
FROM    mtDNA mt
FULL OUTER JOIN Ychr y
ON              mt.TaxaID = y.TaxaID
LEFT JOIN       Cultural c
ON              mt.TaxaID = c.TaxaID
OR              y.TaxaID = c.TaxaID)
```

APPENDIX B: YCHR AND CULTURAL DATASET CHARACTER POSITION LISTS

Table 17: Y-Chromosome STR Loci List, adapted from Butler et al., 2008 [9]

Ychr Dataset Position	**Marker Name**	**Allele Range**	**Repeat Motif(s)**	**GenBank Accession**	**Reference Allele**
1	DYS389_I	9-17	TCTG, TCTA, TCTG, TCTA	AC004617	12
2	DYS389_II	24-34	TCTG, TCTA, TCTG, TCTA	AC004617	29
3	DYS385a	10-19	GAAA	AC022486	11
4	DYS385b	10-19	GAAA	AC022486	11
5	DYS391	6-14	TCTA	AC011302	11
6	DYS390	17-28	TCTA, TCTG	AC011289	24
7	DYS393	9-17	AGAT	AC006152	12
8	DYS392	6-17	TAT	AC011745	13
9	DYS19	10-19	TAGA	AC017019	15
10	DYS437	13-17	TCTA	AC002992	16
11	DYS438	6-14	TTTTC	AC002531	10
12	DYS439	9-14	AGAT	AC002992	13

Table 18: Cultural Ethnographic Atlas Question List, from Gray et al., 1998 [23]

Cultural Dataset Position	**Ethnographic Atlas Codebook**
1	Gathering
2	Hunting
3	Fishing
4	Animal Husbandry
5	Agriculture
6	Mode of Marriage (Primary)
7	Mode of Marriage (Alternate)
8	Domestic Organization
9	Marital Composition: Monogamy and Polygamy
10	Marital Residence with Kin: First Years
11	Transfer of Residence at Marriage: After First Years
12	Marital Residence with Kin: After First Years
13	Marital Residence with Kin: Alternate Form
14	Transfer of Residence at Marriage: Alternate Form
15	Community Marriage Organization
16	Community Marriage Organization
17	Largest Patrilineal Kin Group

18	Largest Patrilineal Exogamous Group
19	Largest Matrilineal Kin Group
20	Largest Matrilineal Exogamous Group
21	Largest Matrilineal Kin Group
22	Secondary Cognatic Kin Group: Kindreds and Ramages
23	Cousin Marriages (Allowed)
24	Subtypes of Cousin Marriages
25	Preferred rather than just Permitted Cousin Marriages
26	Subtypes of Cousin Marriages (Preferred rather than just Permitted)
27	Kin Terms for Cousins
28	Intensity of Agriculture
29	Major Crop Type
30	Settlement Patterns
31	Mean Size of Local Communities
32	Jurisdictional Hierarchy of Local Community
33	Jurisdictional Hierarchy Beyond Local Community
34	High Gods
35	Games
36	Post-partum Sex Taboos
37	Male Genital Mutilations
38	Segregation of Adolescent Boys

39	Animals and Plow Cultivation
40	Predominant Type of Animal Husbandry
41	Milking of Domestic Animals
42	Subsistence Economy
43	Descent: Major Type
44	Sex Differences: Metal Working
45	Sex Differences: Weaving
46	Sex Differences: Leather Working
47	Sex Differences: Pottery Making
48	Sex Differences: Boat Building
49	Sex Differences: House Construction
50	Sex Differences: Gathering
51	Sex Differences: Hunting
52	Sex Differences: Fishing
53	Sex Differences: Animal Husbandry
54	Sex Differences: Agriculture
55	Age or Occupational Specialization: Metal Working
56	Age or Occupational Specialization: Weaving
57	Age or Occupational Specialization: Leather Working
58	Age or Occupational Specialization: Pottery Making
59	Age or Occupational Specialization: Boat Building

60	Age or Occupational Specialization: House Construction
61	Age or Occupational Specialization: Gathering
62	Age or Occupational Specialization: Hunting
63	Age or Occupational Specialization: Fishing
64	Age or Occupational Specialization: Animal Husbandry
65	Age or Occupational Specialization: Agriculture
66	Class Stratification
67	Class Stratification
68	Class Stratification (Endogamy)
69	Class Stratification (Endogamy)
70	Type of Slavery
71	Former Presence of Slavery
72	Succession to the Office of Local Headman
73	Type of Hereditary Succession
74	Inheritance Rule for Real Property (Land)
75	Inheritance Distribution for Real Property (Land)
76	Inheritance Rule for Movable Property
77	Inheritance Distribution for Movable Property
78	Norms of Premarital Sexual Behavior of Girls
79	Prevailing Type of Dwelling: Ground Plan
80	Prevailing Type of Dwelling: Floor Level

81	Prevailing Type of Dwelling: Wall Material
82	Prevailing Type of Dwelling: Shape of Roof
83	Prevailing Type of Dwelling: Roofing Materials
84	Secondary or Alternative House Type: Ground Plan
85	Secondary or Alternative House Type: Floor Level
86	Secondary or Alternative House Type: Wall Material
87	Secondary or Alternative House Type: Shape of Roof
88	Secondary or Alternative House Type: Roofing Materials
89	Region
90	Political Succession for the Local Community
91	Trance States
92	Societal Rigidity

APPENDIX C: DIMENSIONALITY REDUCTION CODE

```
library(tidyr)
library(readr)
library(dplyr)
library(purrr)
library(psych)
library(MASS)
library(dimRed)

#####################################################################
## Input as Table
# Choose Dataset (Inner Join or Outer Join)
CombinedData <- read_csv("DimensionalityReduction_CombinedData_InnerJoin.csv") #Inner Join Data
colnames(CombinedData)[1] <- "TaxaID" #Fix <U+FEFF> Issues
mtDNA_cols <- paste0("mtDNA.pos", seq(1:16590))
Ychr_cols <- paste0("Ychr.pos", seq(1:12))
Cultural_cols <- paste0("Cultural.pos", seq(1:92))

## Separate Strings into Individual Columns
SepCombinedData <- CombinedData %>%
  separate(.,
           mtDNA,
           mtDNA_cols,
           sep = seq(1:12),
           remove = TRUE) %>%
  separate(.,
           Ychr_STR,
           Ychr_cols,
           sep = seq(1:16590),
           remove = TRUE) %>%
  separate(.,
           Cultural_EthnogAtlas,
           Cultural_cols,
           sep = seq(1:92),
           remove = TRUE)

## Remove NA Columns
SepCombinedData <- SepCombinedData[!is.na(names(SepCombinedData))]

## Clean Up Memory to Avoid Error
rm("CombinedData", "Cultural_cols", "mtDNA_cols", "Ychr_cols")
gc()
memory.limit(size=99999)

## Dummy Code Each Column and Write to Disk
DummyCodedCombinedData <- as.data.frame(lapply(SepCombinedData, dummy.code))
```

```
DummyStartLoc <- which(colnames(DummyCodedCombinedData)=="Cultural.pos1.1")

DummyCodedCombinedData <- cbind(SepCombinedData[,1:5],
                                DummyCodedCombinedData[,DummyStartLoc:
                                ncol(DummyCodedCombinedData)]) #Append Data
DummyCodedCombinedData <- DummyCodedCombinedData %>%
  dplyr::select(-starts_with("X")) #Remove Variable With Only 1 Value (Starts with X)

## Clean Up Memory
rm(SepCombinedData)
gc()

## Collapse Data By TaxaID
CollapsedDummyCodedCombinedData <- DummyCodedCombinedData %>%
  dplyr::select(-GuthrieZone,
                -CulturalTaxa,
                -mtDNATaxa,
                -YchrTaxa) %>%
  group_by(TaxaID) %>%
  summarise_all(mean)

rownames(CollapsedDummyCodedCombinedData) <- CollapsedDummyCodedCombinedData$GuthrieZone
CollapsedDummyCodedCombinedData$TaxaID <- NULL
CollapsedDummyCodedCombinedData$GuthrieZone <- NULL
CollapsedDummyCodedCombinedData$CulturalTaxa <- NULL
CollapsedDummyCodedCombinedData$mtDNATaxa <- NULL
CollapsedDummyCodedCombinedData$YchrTaxa <- NULL

## Clean Up Memory
rm(DummyCodedCombinedData)
gc()

#####################################################################
## Multidimensional Scaling
## Non-Metric MDS
MDS_NonMetric_Combined <- isoMDS(dist(CollapsedDummyCodedCombinedData),
                                 y = cmdscale(dist(CollapsedDummyCodedCombinedData),
                                              k = 2),
                                 k = 2,
                                 maxit = 500,
                                 trace = TRUE,
                                 tol = 1e-3,
                                 p = 2)
#Extract Points
MDS_NonMetric_Combined_Points <- as.data.frame(MDS_NonMetric_Combined$points)
#Add Column and Rename
```

```
MDS_NonMetric_Combined_Points$TaxaID <- rownames(CollapsedDummyCodedCombinedData)
colnames(MDS_NonMetric_Combined_Points) <- c("x","y","TaxaID")

############
##Laplacian Eigenmaps

#Find optimal number of k in k-nn
knn_accuracy <- data.frame(k = 0,
                           accuracy = 0)
for (k in 1:nrow(CollapsedDummyCodedCombinedData)){
  knn <- class::knn(CollapsedDummyCodedCombinedData,
                    CollapsedDummyCodedCombinedData,
                    rownames(CollapsedDummyCodedCombinedData),
                    k=2)
  knn_prop <- prop.table(table(knn, rownames(CollapsedDummyCodedCombinedData)))
  accuracy <- sum(diag(knn_prop))/sum(knn_prop)
  iter_accuracy <- data.frame(k = k,
                              accuracy = accuracy)
  knn_accuracy <- rbind(iter_accuracy, knn_accuracy)
}

most_accurate_k <- knn_accuracy[which.max(knn_accuracy$accuracy),]$k

CollapsedDummyCodedCombinedData_dimRed <- dimRedData(CollapsedDummyCodedCombinedData[,
1:ncol(CollapsedDummyCodedCombinedData)]) #Convert to dimRed class
leim <- LaplacianEigenmaps() #S4 Object for LE
LEpars <- list(ndim = 2, #Define Parameter Set for LE
               sparse = "knn",
               knn = most_accurate_k,
               #eps = 0.1,
               t = Inf,
               norm = FALSE)

LE_Combined <- leim@fun(CollapsedDummyCodedCombinedData_dimRed, LEpars)
LE_Combined_Points <- as.data.frame(LE_Combined@data@data) #Extract Points
LE_Combined_Points$TaxaID <- rownames(CollapsedDummyCodedCombinedData) #Add Column and Rename
#colnames(MDS_NonMetric_Combined_Points) <- c("x","y","z","TaxaID")
```

APPENDIX D: MACHINE LEARNING DATA PREPARATION CODE

```
library(tidyr)
library(readr)
library(dplyr)

############
## Load Full Data
CombinedData <- read_csv("DimensionalityReduction_CombinedData_OuterJoin.csv") #Outer Join Data
colnames(CombinedData)[1] <- "TaxaID" #Fix <U+FEFF> Issues
CombinedData[ CombinedData == "NULL" ] <- NA



## Create mtDNA Dataset
mtDNA_cols <- paste0("mtDNA.pos", seq(1:16590))

mtDNAData <- CombinedData %>%
  select(-TaxaID,
         -CulturalTaxa,
         -mtDNATaxa,
         -YchrTaxa,
         -Cultural_EthnogAtlas,
         -Ychr_STR) %>%
  filter(complete.cases(.)) %>%
  separate(.,
           mtDNA,
           mtDNA_cols,
           sep = seq(1:16590),
           remove = TRUE)
# Remove NA Columns
mtDNAData <- mtDNAData[!is.na(names(mtDNAData))]
mtDNAData$GuthrieZone <- as.factor(mtDNAData$GuthrieZone)

## Create Ychr Dataset
Ychr_cols <- paste0("Ychr.pos", seq(1:12))

YchrData <- CombinedData %>%
  select(-TaxaID,
         -CulturalTaxa,
         -mtDNATaxa,
         -YchrTaxa,
         -Cultural_EthnogAtlas,
         -mtDNA) %>%
  filter(complete.cases(.)) %>%
  separate(.,
           Ychr_STR,
           Ychr_cols,
```

```
            sep = seq(1:12),
            remove = TRUE)
# Remove NA Columns
YchrData <- YchrData[!is.na(names(YchrData))]
YchrData$GuthrieZone <- as.factor(YchrData$GuthrieZone)

## Create Genetic Only Dataset
GeneticData <- CombinedData %>%
  select(-TaxaID,
         -CulturalTaxa,
         -mtDNATaxa,
         -YchrTaxa,
         -Cultural_EthnogAtlas) %>%
  filter(complete.cases(.)) %>%
  separate(.,
           mtDNA,
           mtDNA_cols,
           sep = seq(1:16590),
           remove = TRUE) %>%
  separate(.,
           Ychr_STR,
           Ychr_cols,
           sep = seq(1:12),
           remove = TRUE)
# Remove NA Columns
GeneticData <- GeneticData[!is.na(names(GeneticData))]
GeneticData$GuthrieZone <- as.factor(GeneticData$GuthrieZone)

## Create Cultural Datasets
Cultural_cols <- paste0("Cultural.pos", seq(1:92))

CulturalData <- CombinedData %>%
  select(-TaxaID,
         -CulturalTaxa,
         -mtDNATaxa,
         -YchrTaxa,
         -mtDNA,
         -Ychr_STR) %>%
  filter(complete.cases(.)) %>%
  separate(.,
           Cultural_EthnogAtlas,
           Cultural_cols,
           sep = seq(1:92),
           remove = TRUE)
# Remove NA Columns
CulturalData <- CulturalData[!is.na(names(CulturalData))]
```

```
CulturalData$GuthrieZone <- as.factor(CulturalData$GuthrieZone)

## Create Combined Datasets
CombinedData <- CombinedData %>%
  select(-TaxaID,
         -CulturalTaxa,
         -mtDNATaxa,
         -YchrTaxa) %>%
  #filter(complete.cases(.)) %>%
  separate(.,
           mtDNA,
           mtDNA_cols,
           sep = seq(1:16590),
           remove = TRUE) %>%
  separate(.,
           Ychr_STR,
           Ychr_cols,
           sep = seq(1:12),
           remove = TRUE) %>%
  separate(.,
           Cultural_EthnogAtlas,
           Cultural_cols,
           sep = seq(1:92),
           remove = TRUE)
# Remove NA Columns
CombinedData <- CombinedData[!is.na(names(CombinedData))]
CombinedData$GuthrieZone <- as.factor(CombinedData$GuthrieZone)
```

APPENDIX E: LINKS TO RESEARCH MATERIALS

The visualizations generated in this work as hosted online as interactive dashboards on Tableau Public (cford38!/vizhome/BantuMigration/BantuMigration/AnIntegratedPhylogeographicAnalysisoftheBantuMigration).

All scripts and data are available on GitHub (colbyford/BantuMigration).