Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Katherine Bobrek April 9, 2019

Genomic Analysis and Natural Selection Scan of Mexican Mayan and Indigenous Populations

by

Katherine Bobrek

Dr. John Lindo
Adviser

Department of Anthropology

Dr. John Lindo

Adviser

Dr. Craig Hadley

Committee Member

Dr. Helena Pachón

Committee Member

2019

Genomic Analysis and Natural Selection Scan of Mexican Mayan and Indigenous Populations

By

Katherine Bobrek

Dr. John Lindo

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Anthropology

2019

Abstract

Genomic Analysis and Natural Selection Scan of Mexican Mayan and Indigenous Populations
By Katherine Bobrek

In recent years, the field of human population genetics has increased the possibilities of reconstructing human histories. Using modern and ancient DNA, this kind of analysis can complement archaeological and historical accounts of a population history. In this study, human population genetics methods will be applied to a population of Maya and similar indigenous communities in Mexico. The objectives of such study were two-fold. First, the mating patterns and migration history of the Maya, in relation to other groups in the Americas, were reconstructed. Second, a natural selection scan of the Maya population was done to analyze specific markers of natural selection in this population. Having obtained nuclear genomic data for this population of interest from previously published studies, the Mayan nuclear genomes were compared to a set of 55 ancient and modern nuclear reference genomes from North, Central, and South America to answer these objectives. This comparison was done through computational analysis by employing principal component analysis (PCA) and the population branch statistic (PBS). To answer the first objective, the results of the PCA supported the idea that the Maya of Mexico mated with groups in both northern and southern America, creating a spectrum of genetic similarity that correlates to geography. To answer the second objective regarding natural selection, the genes with the highest PBS scores from the natural selection scan were analyzed. Several of these genes were associated with calcium regulation, brain development, olfactory receptors, and the nervous system (RNU5F-1, CDH18, OR52Z1, ROBO2, RAPGEF2, CNTN4, CAMK2D, CRBN). The functional correlations for certain genes are connected to a multitude of functions, including but not limited to fetal adrenal gland and muscle trunk cells; adipose cells; muscle and skeletal tissue; primary T cells in blood; brain cells; neuronal progenitor cells; and mesoderm, ectoderm, and endoderm cultured cells. By analyzing the genes specifically selected for in the Maya population, the history of this population can be reconstructed. Overall, these results broaden our knowledge of the landscape in which the modern Mayan population evolved over their thousands of years of history.

Genomic Analysis and Natural Selection Scan of Mexican Mayan and Indigenous Populations

By

Katherine Bobrek

Dr. John Lindo

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Anthropology

2019

Acknowledgements

I would first like to thank my thesis advisor Dr. John Lindo for all his time and effort in aiding in the analysis and editing of this thesis. I also want to thank my committee members, Drs. Helena Pachón and Craig Hadley and their commitment to my project. Additionally, I would like to thank Dr. Kristin Phillips, and the Department of Anthropology for all their guidance in navigating the thesis process. Lastly, I would like to thank my parents and brother, as well as my friends, for their support throughout this journey.

Table of Contents

Tables

Figures

Viejos, niños, hombres y mujeres, se volvían hormigas después de la cosecha, para acarrear el maíz; hormigas, hormigas, hormigas, hormigas… (Asturias, 1989, pg. 368)

Old folk, young folk, men and women, they all become ants after the harvest, to carry home the maize: ants, ants, ants, ants… (Asturias, 1975, pg. 329)

I. Introduction

The Maya are a dichotomy—a culture with ancient underpinnings, spanning thousands of years of history, and yet, also a modern people alive today. On one hand, the Maya as an ancient civilization are widely known. Despite this, the Mayan picture is far from fully understood. The past and present of the Maya can be jumbled; their past as an ancient civilization is frequently equated as their identity as a modern people. However, the Mayan story is one that encapsulates both. A past born from forests of Central America developed into a diverse, dynamic civilization, later afflicted by collapse, and even later still, upheaval after violent Spanish colonization. The Mayan story continued, and in the modern era, the Maya identity remains, separate from its ancient underpinnings, and yet, a descendant of such past. To truly model this population history over time requires contributions from multiple fields: archaeology, anthropology, history, and linguistics. However, a new field has emerged that can add another dynamic to the Mayan narrative.

Due to technological and biomolecular advancements, the rise of genomics in the past twenty years has enabled the study of both modern and ancient genomes. With this new form of data, groundbreaking discoveries have been made in paleoanthropology, human population genetics, and evolution (Shapiro and Hofreiter 2014). From suggesting mysterious interactions between archaic hominins such as Neanderthals and Denisovans with modern humans, to studying ancient population demography globally, recent work has shown the ability of ancient

and modern DNA to create a fuller picture of human population history and evolution (Slatkin and Racimo 2016).

This new field has not yet fully been applied to the study of Native American population demography. Central America remains an understudied region, and thus, the genomic story of the Maya remains largely untold. While there have been past studies of the Maya, these have used mitochondrial DNA or dental and osteological data (Sochtig et al. 2015). This, unfortunately, has been inadequate in creating a comprehensive population history. However, genomic research can not only create this more comprehensive picture, but also, the impact of such work has a wide range, with "implications for how we comprehend population genetic structure, demographic histories, and even health and disease" (Bolnick et al. 2016). Therefore, this thesis will not only address the lack of data in Mayan genomes, but it will test natural selection using genomic data and computational analysis, allowing for further applications in the future.

Using the modern nuclear DNA of Mayan and other indigenous peoples of Mexico, this research will create a population-specific history for the Maya by the use of two computational analytical tools. First, basic investigation will include principal component analysis (PCA), as a means of comparing genetic similarity between other modern and ancient populations in the Americas. Second, a selection scan will be done for two populations—the Mexican Maya, the focus of this study, and the Andean Aymara, a source for comparison—to evaluate if population-specific natural selection occurred. Additionally, a literature review is added as a comparative tool to evaluate genomic results to other sources of evidence, namely archaeology and history.

With such research, not only will more data be collected from an understudied region, but also, a clearer picture will be illustrated between natural selection and mating patterns amongst the modern Maya today.

II. Literature review

a. Overview of Mayan Civilization

To fully understand the population history that genomic analysis can provide, the current knowledge of the ancient Maya must be explored. Ancient Mayan civilization spanned across Central America, encompassing regions of modern-day Mexico, Belize, Guatemala, El Salvador, and Honduras (Leventhal et al. 2012). The earliest evidence of Mayan civilization is from 1800 BC; however, the earliest inhabitants of the region can be traced back to 9500 BC, otherwise known as the Paleoindian period (Leventhal et al. 2012, McKillop 2006). This period of pre-Mayan civilization was previously not well known until recent genomic analysis, which investigated migration patterns of early humans into North, Central, and South America (Bolnick et al. 2016). The details of such study will be further explored later in this review.

However, Mayan civilization from 1800 BC onwards is well documented in the archaeological record. Historians have separated the timeline of the civilization into five periods: Preclassic, Middle Preclassic, Late Preclassic, Terminal Classic, and Postclassic (Ibarra-Rivera et al. 2008). The details of this timeline, as well as major points, are detailed in Figure 1 below.

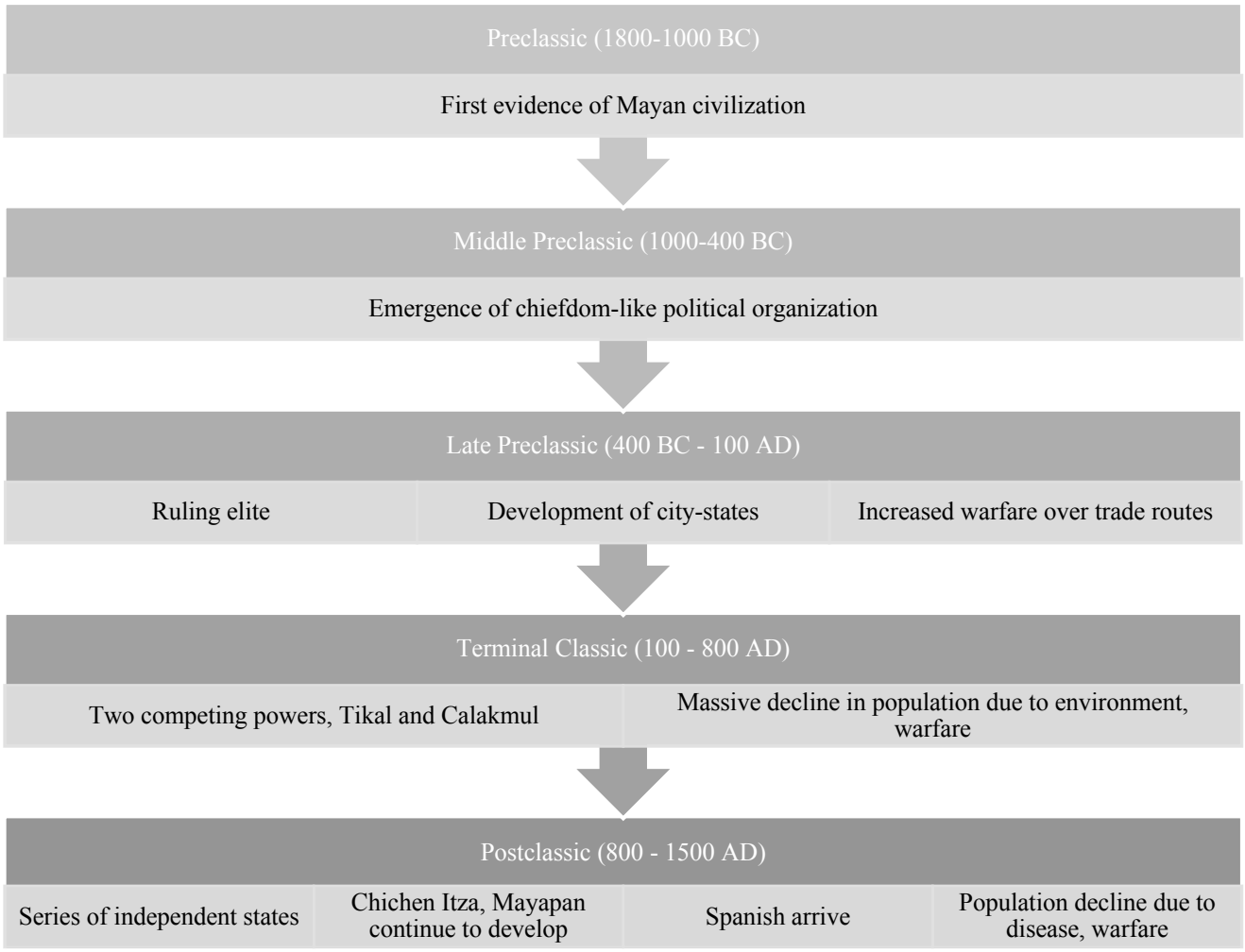

Figure 1. Timeline of ancient Mayan civilization, 1800 BC – 1500 AD.

As has been shown, the transition of one period to the next is marked by rise and fall of the Maya, and by studying these periods, these same motifs can be sought in the Mayan genome. In other words, the changing social structure across periods can affect demographic patterns. For instance, factors such as the fragmented political organization, as well as the irregular distribution of resources, resulted in somewhat scattered settlement patterns, beginning in the Preclassic period (Ford et al. 2009). However, continuing further, the Maya experienced two population collapses, one due to yet unconfirmed causes (but perhaps environmental overexploitation and warfare) in the Terminal Classic period, followed by a second, due to

violent Spanish conquest at the end of the Postclassic period. As this thesis will investigate, these population changes due to environmental, political, and societal effects are hypothesized to have resulted in genetic effects across the population as a whole.

However, these periods in ancient Mayan history have not only affected population structure and genomes, but they have also shaped ancient Mayan culture and language, which has continued into the modern era. By applying these demographic effects of population structure on Mayan culture and language, and specifically exploring how this population history relates to modern Maya today, a more cohesive picture can be drawn between the ancient and modern Maya, an overarching theme of this research. Another dichotomous picture of the Maya emerges when viewing the relationship between population structure with Mayan culture and language. Across the Mayan region, there is a certain cultural continuity; Sochtig et al. (2015) explain, stating: "Mayan art, architecture and rituals suggest a high degree of cohesiveness throughout their domain." Countering this, over twenty-eight languages are spoken by the modern Maya (Ibarra-Rivera et al. 2008). Furthermore, among the Maya, more than twenty distinct ethnic groups exist, including the Mam, Q'eqchi', and K'iche' (Sochtig et al. 2015). Evidently, the idea of a monolithic ancient Mayan state, as well as a monolithic modern Mayan identity, is false; rather, it is a conglomeration of a unique population history.

To elucidate, dynamic population structure can account for the contrast outlined in the previous paragraph between Mayan culture, language, art, and architecture. From linguistic and archaeological evidence, it can be inferred that groups of the Maya lived in isolation across large distances from each other (Canuto et al. 2018). This was due to the ecology and environment of the region, as well as to political structure and subsequent warfare. However, critically, these isolated groups of Mayas did come into contact with one another, through complex trade

networks, long-distance marriage, and captive-taking (Sochtig et al. 2015). While distances between groups were large enough to create linguistic and ethnic differences, mechanisms were in place to allow the spread of ideas between groups. Of interest to this thesis is if this same dissemination occurred genetically as well, as will be tested. Using modern Mayan nuclear DNA, this population history will be tested and compared to current knowledge elucidated from other modes of evidence.

How does the archaeological and historical reconstruction of ancient Mayan population demography connect to modern Maya living today? Research such as the work outlined in this thesis adds depth in understanding the evolving identity of the Maya, which is a product of its ancient population history. To fully understand this connection, the modern Mayan identity will be further explored.

b. Modern Maya today

While ancient Mayan civilization influenced Maya today, modern Maya are also a separate entity, affected by the flux of history and current events. A common misconception of the modern Maya is that they simply do not exist; in other words, the Maya are only a relic of an ancient civilization (Leventhal et al. 2012). This postulate is irrevocably false. Today the Maya constitute the majority of Guatemalans; however, Leventhal et al. (2012) describe their current status, stating: "the Maya have been an underclass: oppressed by the Spanish, other Europeans, Ladinos (individuals with mixed non-indigenous and indigenous ancestry), and even by their own people". While this status has been changing in recent years as Mayans gain more sociopolitical and economic power, the history of the Maya as a subjugated people over the past 100 years is integral to understanding a component of modern Maya identity.

As a case study of the history of modern Maya, and in a larger sense, indigenous peoples in the region, a deeper look at the history of Guatemalan Maya in the 20th century will be briefly reviewed. This history is fraught with "corruption, instability, ethnic discrimination, slow economic development, and systematic violence", due to political coups, decades of military repression, and genocide (Kistler 2014).

An example of the evolving Maya identity in the 20th century is evidenced during the regime of Jorge Ubico, dictator of Guatemala from 1931 to 1944 (Kistler 2014). The Mayan people and culture were objectified as a tool for the Guatemalan state in the form of ethnotourism through Ubico's National Summer Fairs. Little (2008) explains the era, stating: "Mayas who were incorporated into the fair served multiple agendas for Ubico: symbols of Guatemala's past and impediment to progress, tourist attractions, and an underutilized economic resource". This marked the beginning of Mayan culture as a vehicle for progressing Guatemalan economy and politics through tourism, as well as a certain kind of voyeurism.

Further exploitation and discrimination followed in the twentieth century. Shortly after the dictatorship of Ubico, the Guatemalan Civil War began, which lasted from 1960 to 1996, a time during which many Maya faced marginalization in Guatemala. During this period, under the dictatorship of Ríos Montt in 1982, the Maya suffered through genocide, resulting in the deaths of ~86,000 individuals (Kistler 2014). Given these horrific and violent events of the twentieth century, Pan-Mayan activism emerged in the 1970s as a means of fighting the inequality faced by the Maya in Guatemala, through the *clasista* and *culturalista* movements, which successfully fought for class-based equality and a redefinition of Mayan identity, respectively (Kistler 2014). Looking at the status of the Maya today, their identity is thus deeply affected by not only the

atrocities committed against them, but also by the counter-movements led by the Maya to seek equality and the autonomy of their indigenous identity.

It is frequently noted in the literature that before the descent of tourism in the region, the Maya had partially abandoned their indigenous traditions (Medina 2003). This viewpoint is controversial. Some would say that the importance of the ancient Maya is superficial; Leventhal et al. (2012), for example, describes this noting: "One might say that the ancient Maya are celebrated more for the tourism and economic clout they bring to the region, than for the role they play in providing cultural and social continuity to the modern Maya world". However, there is dissenting opinion as well. While violent Spanish colonization forced the Maya to convert to Catholicism, some pre-Columbian religious traditions remain to the modern day. For example, in the highland Chiapas, there is a following devoted to Señor de Tila, a fusion of Jesus Christ and Ik'al, a Mayan deity (Schuster 1997). The relationship the modern Maya have with the ancient Maya is a complex one, and not easily characterized.

However, the assumption that the Maya—as a uniform group—have either lost or retained some of their ancient traditions is a large leap to make. This kind of homogeneity is a complicated viewpoint to defend, and the concept of the modern Mayan identity should be viewed somewhat differently; Fischer (1999) explains that Mayan culture can be considered "As a historically continuous construction that adapts to changing circumstances while remaining true to a perceived essence of Mayaness". When considering the Maya as a people constantly adapting to change, it is important to examine the most recent of metamorphoses. Many Maya live in diaspora communities in the United States, particularly in Southern California, San Francisco, Houston, Miami, and Chicago (Leventhal et al. 2012). Those living in these communities are confronted with the difficulties of retaining their Mayan language, culture, and

spirituality in the face of Euro-American culture (Batz 2014). It is clear that beginning with Spanish colonialism, the Maya have been faced with challenges to identity that pervade into the modern day.

In short, looking at the sum of the ancient Maya, their modern Mayan descendants and culture, politics, and identity, it is clear that the archetypal image of the Maya is too one-dimensional. Castaneda (2004) encapsulates this idea, explaining:

> *Too often the public eyes of the international media and academic community assimilate all Maya to a homogenizing category of a uniform identity. Many have noted the way archaeological and touristic discourses construct an image of the Maya as mysterious and living outside of time. Similarly, the discourses that celebrate the Maya as a culture and people surviving oppression, modernity, and capitalism through struggles against the national (and racialist) elite, create a monolithic stereotype that erases the heterogeneity and cultural diversity of the Maya.*

By assessing this concept of an expressly heterogeneous people, an image of the modern Maya is formed that comes closer to the truth. It is one of a deeply diverse background, a diversity that has been in place since the very beginnings of ancient Mayan civilization. From the very first city-states in the Preclassic period to the millions of Mayas inhabiting the Americas today, the only invariable statement that can be made is that the Maya never were invariable at all. Thus, when reconstructing their population history and analyzing their genomes for natural selection, as is the goal of this thesis, how this demography fits in to this complex, modern identity is critical to understanding the history of the Maya as a whole, and genetic effects of this can affect the biology and health of the modern Maya.

c. Human population genetics

To reconstruct such a complex population history presents an exciting challenge. To date, nuclear DNA analysis of the Maya has been sparse. There has been work with the mitochondrial genome (mtDNA); however, it is widely accepted that mtDNA studies do not account for a full picture of human population history (Bolnick et al. 2016). Mitochondrial DNA is passed down the maternal line, so when looking at mtDNA by itself, the paternal line is missing. This can cause biases in analysis. Nuclear DNA presents a fuller picture of both lineages. The full extent of the work done on Mayan mtDNA and the subsequent shortcomings in such analysis will be discussed later in this work. It is imperative that nuclear DNA be studied instead. However, it is vital for another reason, as well. Archaeological excavation in the Mayan region remains difficult due to the dense jungle (Ford et al. 2009). Population genetics can allow for a new resource in filling in gaps in knowledge.

The broad goal of population genetics is "to infer the past history of populations and describe the evolutionary forces that have shaped their genetic variation" (Tataru 2016). While the field of population genetics is not a new one, having developed in the 20th century, recent breakthroughs have caused a renaissance of new possibilities, creating a shift from theoretical to more data-oriented research (Nielsen and Slatkin 2013). The use of whole nuclear genomes, next-generation sequencing (NGS), Big Data, and increased computational power and methods of analysis are factors which have contributed to the study of genomics, or the genome as a whole. Regardless, main principles of evolution and genetics remain, which will be explored further as a foundation.

Population genetics as a field studies variations in the frequencies of alleles and genotypes (Nielsen and Slatkin 2013). The frequency of an allele is calculated as "the number of

copies of the allele in the population divided by the total number of gene copies in the population", defined below, given N as population size, A as the allele studied, and f_A as the frequency of such allele (Nielsen and Slatkin 2013):

$$f_A = \frac{N_A}{2N}.$$

This fraction represents the population size carrying allele A over the total population. Given this, changes to these frequencies are analyzed to answer questions regarding population change and evolution.

One of the classical frameworks for understanding changes in allele frequencies over time is the Wright-Fisher Model (Tataru et al. 2016). The Wright-Fisher Model, first published in 1930, allows for evolutionary and demographic inference based on allele frequency data; Tataru et al. (2016) states: "The Wright–Fisher model… explicitly accounts for the effects of various evolutionary forces—random genetic drift, mutation, selection—on allele frequencies over time." The Wright-Fisher model can be explained mathematically as a discrete Markov chain (Dat Tran et al. 2013). The basic model, assuming a diploid individual, describes probability P, as follows:

$$P(Y_{N+1} = j | Y_N = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$

for $i, j = 0, \ldots, 2N$

To summarize, for a discrete time step $t + 1$, to create generation $N + 1$, each allele from parent i and parent j can be chosen from generation N. While Wright-Fisher provides the basis for understanding population genetics, the model is built from assumptions—no migration and random mating, to name a few—that do not reflect populations in reality. It is worthwhile to further explore other classical models of population genetics, as others have been developed to model organisms more realistically.

An example of a more realistic model is the coalescent model. A key difference between the Wright-Fisher model and the coalescent model is that while the Wright-Fisher predicts forward in time, the coalescent is a mathematical framework that simulates genealogical relationships backward in time (Orlando and Cooper 2016). This concept of simulating backward in time is illustrated in Figure 2 below. T_n represents discrete time steps which are exponentially distributed. When $n = 5$, T_n denotes the present day. For each time step, $n - 1$, separate lineages fuse, as is denoted by each orange dotted line in Figure 2. At T_1, the coalescent has converged to the most recent common ancestor of the modern populations at T_5.

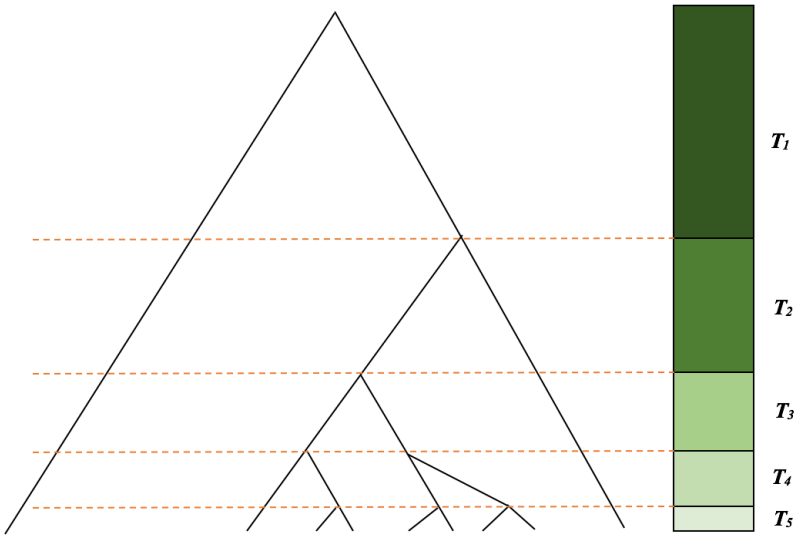

Figure 2. The coalescent model describing branching of a given population over time.

Another critical method in population genetics is natural selection scanning. To gauge the effects of cultural and genetic adaptations at the genomic level, the genome can be scanned to produce genes of interest, which can be correlated to phenotypes, such as altitude adaptation or response to hypoxia (Yi et al. 2010). The population branch statistic (PBS) is an effective method of detecting natural selection. PBS is a measure of distance between populations using

the F_{st} statistic. High scores indicate the possibility of selection (Yi et al. 2010). In this thesis, PBS score will be used to do similar analysis.

Despite the lack of Mayan nuclear DNA, these principles of population genetics have been applied to the Maya as a means of discerning Mayan population history and demography, albeit with mitochondrial DNA (Sochtig et al. 2015).

d. *Previous genomic work in Mayan population history and demography*

Current research on the population history of the Americas has centered greatly on the arrival of humans in the region. This review will begin from this time period and investigate the main ideas identified from current work on Mayan demography, with the knowledge that much of the work with nuclear DNA pertains only to the very earliest of Mayan history.

In terms of early Native American history, most work on Native American population dynamics has been based on archaeological data. However, as genomic data increase, our understanding of the population history of early Native American demography becomes much richer. The archaeological evidence will first be briefly explored, followed by an overview of the genomic evidence of early Native American demography. Lastly, this review will examine the current work later in time, pertaining specifically to Mayan demography.

An important reconstruction made in Native American demography is the migration of humans into the Americas. According to archaeological evidence, North America was colonized at around 13,000 years ago, by those associated with Clovis technology; however, some evidence in both North and South America alludes to even earlier settlement (Slaktin and Racimo 2016). As genomic evidence builds to complement this archaeological evidence, it is coming from two sources: ancient DNA and modern DNA. For example, the 24,000-year-old fossil from Mal'ta in Siberia has yielded ancient nuclear DNA, evidence which helps to illustrate the earliest

beginnings of Native Americans (Slatkin and Racimo 2016). Based on this piece of evidence and others, it can be said that humans first entered the Americas through the Bering Land Bridge from Siberia during the Last Glacial Maximum (28,000-18,000 years ago), after an 8,000-year period of inhabiting Beringia (Bolnick et al. 2016; Raghavan et al. 2015). From there, humans relatively quickly populated the Americas, reaching southern South America by 14,600 years ago. Genomic evidence alludes to a differentiation of the founding population into two groups, happening roughly 13,000 years ago (Bolnick et al. 2016; Raghavan et al. 2015). Even given this population separation, while genetic differences exist across the north to south gradient in the Americas, greater genetic similarity exists in Native Americans, as opposed to other groups globally; Bolnick et al. (2016) explain, stating: "Native Americans exhibit less genomic diversity and more LD [linkage disequilibrium] than do other worldwide populations… within the Americas, genomic diversity decreases as one moves south, whereas the degree of genetic differentiation between populations generally increases." This trend of southern geography and genetic variation could be reflected in the Mayan genome, as genomic distinctions exist between northern and southern Native Americans, with the Maya being classified as central Americans. This genetic difference between different regions in the Americas is further validated by another source of ancient nuclear DNA from the fossil Anzick-1. The fossil, dated to 12,600 years ago and associated with Clovis culture, yielded a DNA sequence in which it was clear that it was either a direct ancestor or a very close relative to a direct ancestor of modern Native Americans (Slatkin and Racimo 2016). Furthermore, Anzick-1 is more closely related to Central and South Americans than to native North Americans; this a fascinating piece of archaeology, given that the fossil was found in modern-day Montana, USA (Bolnick et al. 2016; Posth et al. 2018). Finally, recent research has shown that while the Anzick-1 genome does show continuity

between North American Clovis culture and Central and South Americans, the majority of ancestry of Central and South Americans later in time does not have any Anzick-1 ancestry, alluding to a population structure that was not homogenous (Posth et al. 2018). Looking at all these pieces of evidence as a whole, certain conclusions can be made regarding the ancestors of the Maya and their earliest history. First, the timing of the first humans into the Americas is known to have occurred around 23,000 years ago. Second, these humans spread throughout the Americas, eventually differentiating into different groups at around 13-14,000 years ago. From here, further technological, sociopolitical, and environmental factors would contribute to the evolution of the Mayan group.

After this stage in the Mayan story, nuclear genomic evidence becomes sparse, and most inferences from this point forward are supported by only mitochondrial DNA. Mitochondrial DNA (mtDNA), however, is insufficient to provide a full picture of demographic history, given that it only reveals demographic details of the maternal line. Conversely, nuclear DNA is the sum of both paternal and maternal lines, creating a more accurate picture of the genetic history of an individual. When looking at studies which use mtDNA of modern Maya, the inference is perhaps skewed, and thus, the results should be viewed with some apprehension.

Another key area of research focus has been on the development of agriculture and its effects on the genome. Specifically, the introduction of maize, courgette, bean, and pepper agriculture at around 7000 BC had a significant effect on Mayan, and in a broader sense, Native American, genetics (Gonzalez-Martin et al. 2015; Ochoa-Lugo et al. 2016; Bolnick et al. 2016). As an effect of the growth of agriculture, the subsequent population growth, increase in trade, and development of civilizations such as the Olmec and Maya contributed to an increase in regional genetic homogeneity, according to the mtDNA (Bolnick et al. 2016). Thus, as

populations grew, yet also became more sedentary due to the shift from hunter-gatherer to agricultural societies, regional groups became more genetically homogenous. However, this should not be taken to mean that Mayan populations became completely isolated from one another. Rather, with these shifts in societal structure came increased trade and warfare, and this in turn, led to increased gene flow between groups, with the highest flow during the Preclassic and Classic periods, again according to mtDNA (Sochtig et al. 2015).

Looking between modern Mayans and non-Mayans, specifically those found in Guatemala, adds another facet to the reconstruction of population demography. Comparing the mtDNA of Mayans to non-Mayan mtDNA reveals that despite gene flow, Mayan populations show less genetic diversity than in non-Mayans (Sochtig et al. 2015). This is supported by decreased levels of heterozygosity and can be explained by bottlenecks, founder effects—reduced genetic diversity resulting from a population descended from a small group of colonizing ancestors—population size, and inbreeding—mating of individuals closely related to one another (Sochtig et al 2015; Ibarra-Rivera et al. 2008; Gonzalez-Martin et al. 2015). This evidence of limited genetic heterozygosity complements the archaeological and historical evidence, which supports the hypothesis that the Mayan civilization experienced sharp declines in population (in other words, a bottleneck) during the Terminal Classic period, and again, during Spanish colonization.

The effects of Spanish colonization on Native American—and specifically for the purposes of this review, Mayan—populations have been often described, and yet, the scale of such a catastrophic event is still difficult to imagine. Colonialism had wide-reaching effects; Bolnick et al. (2016) summarize some of them, stating: “Native Americans experienced repeated disease epidemics, warfare, slavery, relocations, changes in community structures, admixture

with non-Native peoples, and the destruction of traditional lifestyles." Naturally, all of these effects would have had led to genetic effects in the mtDNA as well. In particular, disease, warfare, and slavery led to a dramatic population bottleneck (Sochtig et al. 2015; Bolnick et al. 2016; Oland and Palka 2016). In the Yucatan peninsula alone, the Mayan population fell from an estimated 800,000 to 140,000, a population decrease of 83% (Oland and Palka 2016). Given this colossal damage to Mayan population size and structure, a great bottleneck occurred in population demography during this period. Looking further, more effects from Spanish colonization can be found in Mayan genomes. For example, a comparison between Mayan and non-Mayan (Ladino) peoples in South America reveals the differential degree of sex bias between these two populations. Unlike the Maya, Ladino populations are more admixed, having more genetic ancestry from European colonizers. It is widely cited in the literature that during the admixture that took place during European colonization, sex bias occurred (Bryc et al. 2010; Bolnick et al. 2016; Sochtig et al. 2015). In other words, European men contributed a higher proportion of genetic ancestry to the admixed population—in this case, the Guatemalan Ladinos—than Native American men or European women. Thus, in Ladino genomes, it is common to see a higher proportion of Native American ancestry on the X chromosome than on the Y chromosome (Sochtig et al. 2015). However, this trend does not hold for modern Maya populations, who must be the descendants of ancient Maya who came into less contact with Spanish colonizers than other groups of Maya (Oland and Palka 2016). It is clear that the effect of Spanish colonization on the ancient Maya led to modern populations having a differential degree of Mayan ancestry in the genomes of indigenous and admixed populations, as a result of sex bias.

The Mayan population demography, which has just been outlined, has been the sum of knowledge elucidated from genomic analysis. However, looking at the evidence of Mayan population demography post-migration into the Americas, as has been said, all of it has been from mitochondrial DNA. Thus, the current body of work must be expanded, as this thesis hopes to accomplish. These demographic events—migration and mating patterns—will be studied as a part of this thesis, the results of which can thus be compared to previous findings from mtDNA.

It is worth asking what the broader impact is of reconstructing Mayan population history. Why does it matter, and to whom should it matter? Such work has multiple applications, particularly health-based ones in predicting and preventing disease. Going even further past practical applications, what are the cultural implications of genomics and population genetics in relation to Native Americans and their cultures and genomes? With the rise of this field, while its novel uses and the solutions that it provides have been often discussed, another integral piece to this phenomenon is the problems it creates. These solutions, as well as these problems, will be discussed further.

e. Broader applications and cultural implications

The impact of genomic research is gradually being understood by both scientific researchers and the general public. Particularly, its applications in health-based solutions are an increasingly common sub-field of research. From studying racial survival disparity of breast cancer among American women, to overhauling the Finnish healthcare system's statistical evaluation of Big Data, to creating international genomic research initiatives, the breadth of applications in population genetics and genomics in public health is expanding (Daly and Olopade 2015; Palotie et al 2017; Adoga et al. 2014). These same themes can be applied to the Central American population at the focus of this thesis.

In fact, recent work with Mayan populations in Mexico can be used as a case study of this. Mexicans with increased Native American ancestry have an increased risk of diabetes, and Mayans in Mexico are primarily the group with the highest Native American ancestry (Sanchez-Pozos et al. 2018). In Sanchez-Pozos et al's study, specific variants were identified through whole-exome—the portion of the genome containing information coding for proteins—sequencing of the nuclear DNA of twelve modern Mayan individuals. The variants found in genes with connections to diabetes, lipid metabolism, and insulin resistance are described in Table 1 below:

Variant	Gene	Disease association
rs1799999	*PPP1R3A*	Risk of type II diabetes, increased values of HOMA-IR[1]
rs1801702	*APOB*	Total cholesterol and LDL-C[2] in normoglycemic individuals
rs9624	*TPPP2*	HOMA-IR[1]
rs3732083	*GPR1*	Total cholesterol and triglycerides in diabetic individuals

1. Homeostasis model assessment insulin resistance index
2. Low density lipoprotein cholesterol

Table 1. Variant associations with disease among twelve modern Mayan individuals (Sanchez-Pozos et al. 2018).

Given these results, a clearer connection among ancestry, genetics, and disease can be formulated. While this type of research is still developing, recent work has elucidated upon its Native American-specific clinical implications; Harris et al. (2018) expand upon this, stating:

> *The low genetic diversity estimates of Native Americans suggest that there may be an enrichment for rare diseases in Native American ancestry communities… Due to the underrepresentation of Native American ancestry in genomic databases, we hypothesize that Native American communities may have an increase in recessive disease alleles that are unobserved in current clinical databases.*

What recent findings have shown is that not only are there population-specific health consequences encoded within the genome, but they have also shown that Native Americans are underrepresented in these genomic analyses.

While it can be said that Native American ancestry is underrepresented in genomic databases, it is also apparent that among both scientific and wider communities, there is a heightened awareness of the use of DNA in determining ancestry based on past population dynamics, specifically in Native American ancestry in some cases. As genomic technologies have advanced in the past decade, from academic papers to at-home DNA kits, much more focus has been placed on DNA as the final determinant in deciding identity. How does this "genetic fetishism" affect identity, specifically, in terms of the Mayan identity (Tallbear 2013)? This review's earlier exploration of Mayan identity illustrated a complex and longstanding history. Now, over the past decade, some argue that the field of genomics that has emerged once again questions Mayan identity.

In particular, the emergence of Native American DNA analysis in reconstructing population histories brings to light biases in genomics-driven inferences regarding identity. To understand the implications of genomics on Mayan identity, a further investigation must be made into the history of science, culture, and race over the past century. Stemming from the first half of the twentieth century, the relationship between science and race had been more closely intertwined, embodied in the infamous eugenics movement (Wade 2017). With the after-effects of the Nazism of World War II, as well as the work of anthropologists Franz Boas and Claude Levi-Strauss, the study of human populations separated from the race science of years prior (Wade 2017). However, given this fraught foundation, the field has struggled with the

relationship among biology, culture, and society, specifically in the conceptualization of human diversity and the definition of a "population".

In short, how does one quantify human diversity? Currently, this definition is a blend of biological, geographic, and cultural differences. However, this designation can be problematic, especially when describing Native American populations; Tallbear (2013) explains this difficulty, stating: "The populations… that are identified and studied mirror the cultural, racial, ethnic, national, and tribal understandings of the humans who study them… For and by whom are such categories defined?" In other words, this brings to light the inherently biased positionality of those who study genomics; geneticists are often outsiders to the societies which they study, as is the author of this study. From creating population designations to creating inferences regarding these populations, a scientific bias exists that must be recognized and considered when making further conclusions on Mayan identity.

For example, the recent meteoric rise of genomic technologies has led to a disharmonious relationship with the anthropological questions these technologies mean to answer; Tallbear (2013) elaborates, stating: "Anyone—regardless of how much anthropology they know—can do anthropology. This has produced research interpretations that are… strikingly culturally naïve." As the study of human demography through genomics aims to make inferences regarding identity, and as these inferences are made by scientists that are nearly always outsiders to the Native American populations studied, this prescribed identity can further complicate a complex history of Native American identity.

How exactly can these biases affect Native American, and by extension, Mayan identity? Since colonialism, Native Americans have grappled with autonomy and identity, as evidenced in the history of the modern Maya discussed earlier. Now, as genomics researchers prescribe

relationships and connections between peoples, some worry that these conclusions will supersede previous markers of identity. Tallbear (2013) expands upon this question, stating: "The question is, as genetic identities and historical narratives command increasing attention in society, will they come to rival as legitimate grounds for identity claims the existing historical-legal foundations of indigenous governance authority?" In other words, sometimes DNA analysis can be seen as the true determinant, the last word in defining the conclusion to a question. If this genomic analysis is founded on biased analysis—analysis done from an outsider's point of view—and this analysis is taken as hierarchically superior to other determinants of identity, then this could detrimentally alter conceptions of Native American identity.

To reconcile these facts with the research presented in this thesis, the positionality of the researcher should be kept in mind. Furthermore, the results of such research must be holistically viewed in the larger context of past work in Mayan archaeology, linguistics, and politics. To truly present a population demography that answers anthropological questions respectfully, the larger context of that population's culture must be kept in mind.

The goal of this thesis is to collect more data from an understudied region and to create more connections between ancient and modern populations, natural selection, and mating patterns amongst the Maya through the use of computational analysis. However, what must not be lost is the heart of the matter; in the end, it is simply another narrative, added onto thousands of others, to answer a question with no one answer: Who are the Maya?

III. Methods

a. Objectives

To reiterate the objectives outlined in the introduction, the goal of this study is two-fold.

First, by means of principal component analysis (PCA), a fuller picture of genetic similarity amongst the Maya and other Native American groups will be explored. These similarities can lead to interpretations regarding mating patterns and migrations in the Mayan population. Second, by means of a natural selection scan, potential genes under selection in the Mayan population can be examined.

b. Reference genomes

A collection of reference genomes, both modern and ancient, was gathered from the Simons Genome Diversity Project (SGDP) among other sources (Moreno-Mayar et al. 2018; Lindo et al. 2018; Mallick et al. 2016). The non-reference, genomes of interest—the modern Mexican indigenous populations, including the Maya—were obtained through the permission of the authors of Romero-Hidalgo et al. 2017, who were contacted via email. The genomic data were originally collected (via saliva samples) and sequenced by this group. The individuals chosen for nuclear DNA collection identified as Native American—in other words, having parents and grandparents who identified as Native American, being born in Native American communities, and speaking the native language (Romero-Hidalgo et al. 2017). After sequencing, it was determined that ten of the twelve Mexican genomes had 98% indigenous ancestry, save for the two of the twelve genomes, which had 91% (Romero-Hidalgo et al. 2017). With their agreement, their data—in the form of variant call format (VCF) files for each individual in their study—was shared for this project. The SGDP is a publicly accessible database of a multitude of human genomes, and the genomes of interest were thus downloaded from this database. Several ancient genomes such as the Sumidouro Cave genomes and the Upper Sun River Valley genome are also publically accessible through their respective publications (Mallick et al. 2018). Ancient Andean and Aymara genomes were provided by Dr. John Lindo at Emory University from his

previous research (Lindo et al. 2018). These genomes were chosen to provide a diverse picture of Native American populations through time and space. There are, therefore, not only genomes from North, Central, and South America, but there are also ancient and modern genomes. This was done to adequately compare the Mayans of interest to as many relevant populations as possible.

This research obtained IRB approval from Emory. No lab work was done, as all genomic data was downloaded for each individual from the databases outlined above. The data was originally downloaded as a binary alignment map (BAM) file for each individual. The BAM file contained the full genome of each individual. For later analyses, each file was converted to VCF format. In the file for each individual, the alleles for each variant are listed and additionally marked with chromosome location. Then, each VCF containing the genomic information for each individual was merged to create one VCF file containing all genomes listed in Table 2. All genomes are referenced to the hg19/b37 build—the reference assembly of a species used when sequencing a new genome—and are nuclear DNA. The Romero-Hidalgo et al. genomes contain twelve indigenous individuals from Mexico, two of which are Mayan. Ten out of twelve individuals show higher than 95% indigenous background (Romero-Hidalgo et al. 2017). The individuals from Romero-Hidalgo et al. were later used in downstream analyses as the population of interest, the Mexican Maya. For the PCA, all twelve were used. For the natural selection scan, the two genomes with lower indigenous ancestry were excluded. This information has been compiled in Table 2 below. The column labeled "Code name" is identifying information later used in the PCA analysis in the results section.

Population (n)	Code name	Region	Source	Ancient/Modern
Chane (1)	Chane	South America	SGDP[1]	M
Surui (2)	Surui	Brazil	SGDP	M

Karitiana (3)	Karitiana	Brazil	SGDP	M
Piapoco (2)	Piapoco	South America	SGDP	M
Pima (2)	Pima	North America	SGDP	M
Mayan (2)	Maya_Simons	Mesoamerica	SGDP	M
Zapotec (2)	Zapotec	Mesoamerica	SGDP	M
General Mesoamerican (5)	Mixe/Mixtec	Mesoamerica	SGDP	M
Quechua (3)	Quechua	Andes	SGDP	M
Maya (2)	Maya_RH	Mesoamerica	Romero-Hidalgo et al. 2017	M
Nahua (2)	NAH_RH	Mesoamerica	Romero-Hidalgo et al. 2017	M
Tarahumara (2)	TAR_RH	Mesoamerica	Romero-Hidalgo et al. 2017	M
Tepehuano (2)	TEP_RH	Mesoamerica	Romero-Hidalgo et al. 2017	M
Totonaca (2)	TOT_RH	Mesoamerica	Romero-Hidalgo et al. 2017	M
Zapoteca (2)	ZAP_RH	Mesoamerica	Romero-Hidalgo et al. 2017	M
Spirit Cave (1)	Spirit Cave	North America	Mallick et al. 2018	A
Upper Sun River Valley (1)	USR	North America	Mallick et al. 2018	A
Sumidouro Cave (2)	Sumidouro	Brazil	Mallick et al. 2018	A
Ancient Andean (5)	Anc_Andean	Andes	Lindo et al. 2018	A
Aymara (24)	Aymara	Andes	Lindo et al. 2018	M

1. Simons Genome Diversity Project

Table 2. List of all populations used in downstream analyses.

Additionally, all other reference populations were later used in downstream analyses as comparison.

c. Analyses

Preparing genomic data required the use of *SAMtools* and *BCFtools* to properly format BAM data files to VCF data files for analysis (Li et al. 2009; Li 2011). *VCFtools* was used for later manipulation and merging of data files (Danecek et al. 2011).

The first analysis done was a principal component analysis. An R script was used to create the plot with libraries *gdsfmt*, *SNPRelate*, and *ggplot2* (Zheng et al. 2012; Wickham 2012). A PCA uses an orthogonal transformation to convert a group of observations into principle components. These principle components can be plotted along x- and y-axes, with a data point for each individual observation—which in this case is a single genome. Distances between observations correlate to relatedness between observations. The goal of such analysis was to determine inter-relatedness between Native American populations.

Final analysis of the natural selection scan was done with *VCFtools*, followed by an R script to visualize the data in the form of a Manhattan plot, using library *qqman* (Turner 2014). The natural selection scan was done twice, once with the Mexican Maya (N = 10) as the target population and a second time with the Aymara as the target population. The Aymara were analyzed as a type of comparative control. The top hundred hits of the natural selection scan were analyzed using Genecards and HaploReg to ascertain phenotypes associated with genes. This is an outlier method, so these top one hundred hits represent signals above the 0.01% threshold.

IV. Results

a. Principal Component Analysis (PCA)

To fulfill the goal of this thesis to understand mating patterns amongst Native Americans, the following PCA in Figure 3 below reveals the inter-relatedness between reference

populations and the Mexican Maya of interest, as outlined in Table 2. Full nuclear genomes were used. The code names used in the legend to the right of Figure 4 are described fully in Table 2.

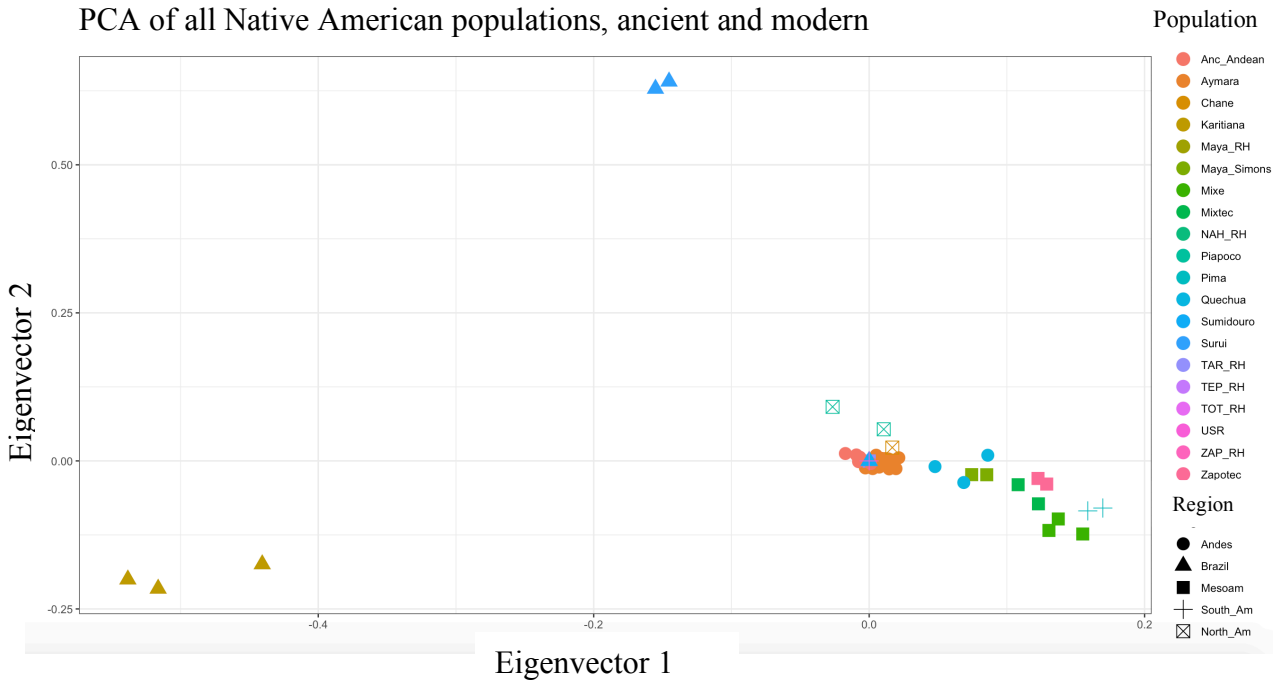

Figure 3. PCA of reference populations

The populations from North America, South America, and Mesoamerica follow a spectrum. North American populations such as the Pima and Upper Sun River Valley (USR) are at the far end of the spectrum, interestingly, closely associated with the Aymara, Quechua, and ancient Andeans. In the middle of the spectrum are the Mesoamericans, whom are more closely associated with South American populations such as the Piapoco. Both modern and ancient Brazilians not only do not follow this spectrum, but they also are different from each other.

b. Natural Selection Scan

To fulfill the goal of this thesis to understand if natural selection occurred amongst the Mexican Maya, the results of the selection scan were plotted below for both targets, Aymara and Mexican Maya. Again, a selection scan of the Aymara was done as a control. The y-axis is the PBS score. Each data point on the Manhattan plot represents a variant along a chromosome.

For the Manhattan plot of the Maya in Figure 4, across all chromosomes there seem to be variants with a higher degree of selection. From the Manhattan plot of the Aymara in Figure 5, it is evident that in the Aymara population, chromosomes 4, 12, 14, and 17 had the variants with the highest degree of selection. In a more general sense, it is clear that selection did occur on the outliers in both data sets, given that they have relatively higher PBS scores than all other variants. Therefore, the top one hundred hits of the scan were further analyzed below for genes of interest.

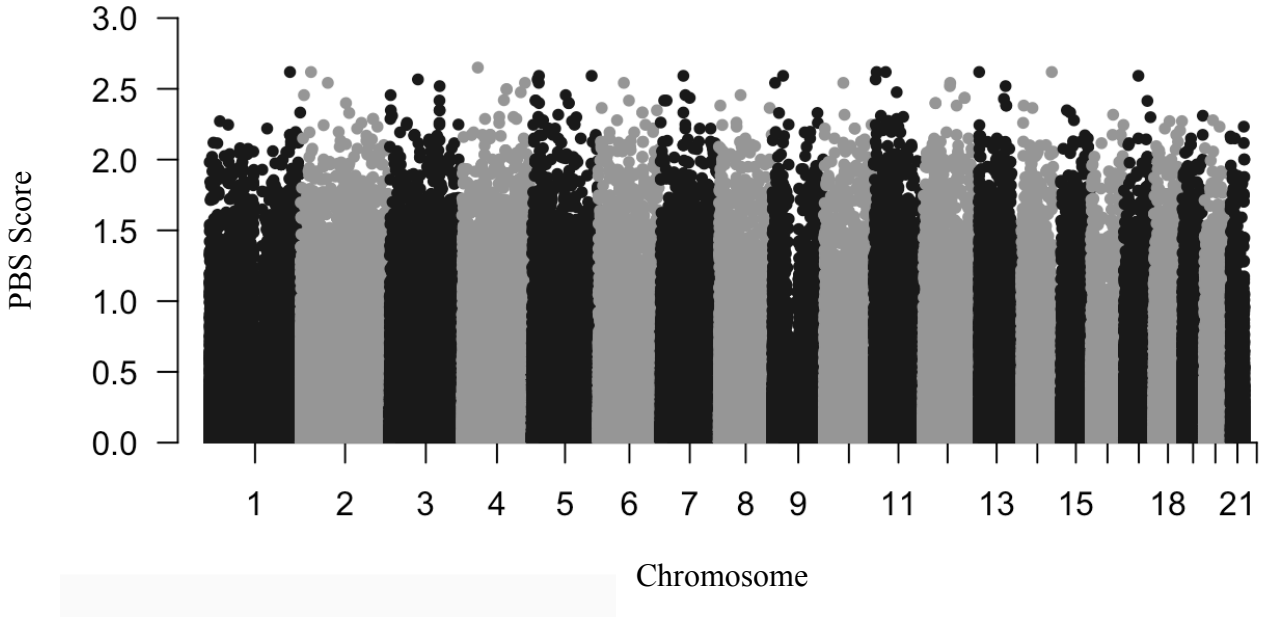

Figure 4. Manhattan plot of Population Branch Statistic (PBS) scores with Mexican Maya as the target population

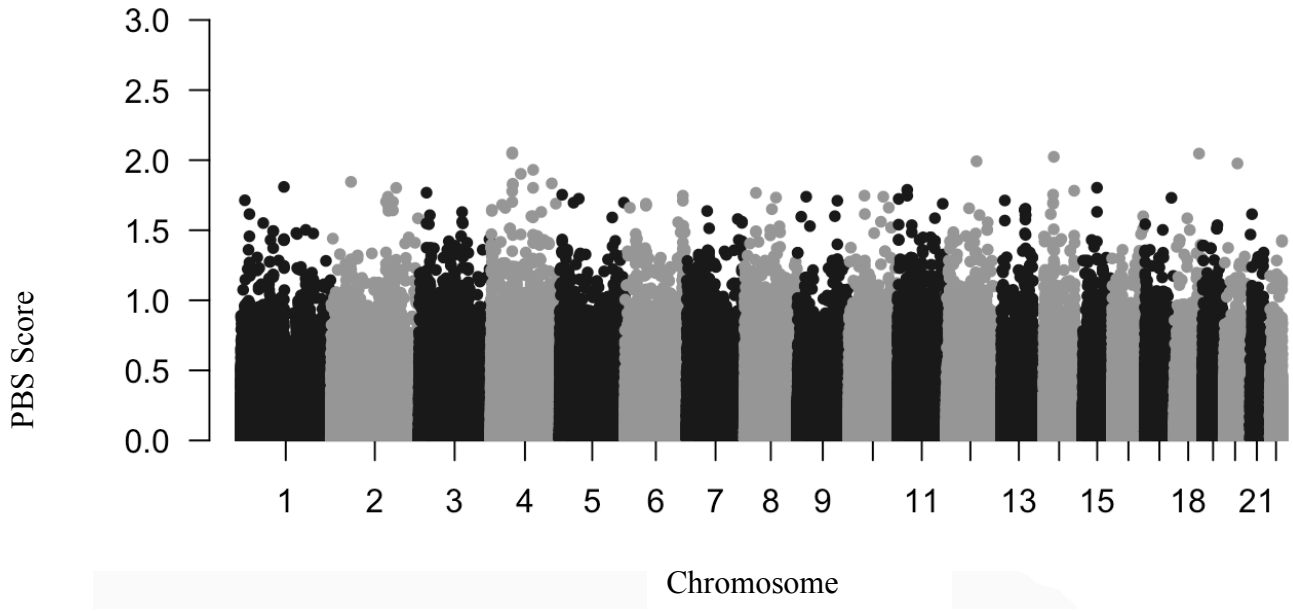

Figure 5. Manhattan plot of Population Branch Statistic (PBS) scores with Aymara as the target population

The results of the natural scan were sorted by population branch statistic (PBS) score. High PBS score indicates possibility of natural selection. Genes with phenotypes of interest are outlined below for both Mexican Maya and Aymara selection scans.

Gene	PBS Score	Description[1]	Phenotype[1]	Functional correlation[1]
GRXCR1	2.65	Glutaredoxin and Cysteine Rich Domain Containing 1	Associated with deafness	N/A[2]
CCDC73	2.62	Coiled-Coil Domain Containing 73	Associated with cryporchidism	N/A
OR10A3	2.62	Olfactory Receptor Family 10 Subfamily A Member 3	Associated with olfactory receptors; transmission of odorant signals; neurotransmitters	N/A
RNU5F-1	2.62	RNA, U5F Small Nuclear 1	Associated with leukodystrophy, development of white matter	Colon cells
WDR20	2.62	WD Repeat Domain 20	Distribution of red blood cell width, metabolism of proteins	Brain hippocampus, adipose, fetal adrenal gland, skeletal and

Gene	PBS Score	Description[1]	Phenotype[1]	Functional correlation[1]
				muscle myoblast cells
KRTAP9-9	2.59	Keratin Associated Protein 9-9	Involved in developmental biology	N/A
MAT2B	2.59	Methionine Adenosyltransferase 2B	Involved in metabolism; variants in these gene associated with changes to hemoglobin, carotid arteries	N/A
CDH18	2.59	Cadherin 18	Expressed specifically in the central nervous system and is putatively involved in synaptic adhesion, axon outgrowth and guidance	N/A
OR52Z1	2.57	Olfactory Receptor Family 52 Subfamily Z Member 1	Associated with olfactory receptors; transmission of odorant signals; neurotransmitters	N/A
ROBO2	2.57	Roundabout Guidance Receptor 2	Associated with functions in axon guidance and cell migration	ES-WA7 cells
ASAH2B	2.54	N-Acylsphingosine Amidohydrolase 2B	Associated with corneal dystrophy, posterior amorphous, and Alzheimer Disease	N/A
TRHDE	2.52	Thyrotropin Releasing Hormone Degrading Enzyme	Inactivates the neuropeptide thyrotropin-releasing hormone	N/A
RAPGEF2	2.48	Rap Guanine Nucleotide Exchange Factor 2	Involved in neuron migration and in the formation of the major forebrain fiber connections forming the corpus callosum, the anterior commissure and the hippocampal commissure during brain development	N/A
CNTN4	2.46	Contactin 4	Play a role in the formation of axon connections in the developing nervous system; mutations in gene associated with autism	Endoderm cultured cells

Gene	PBS Score	Description[1]	Phenotype[1]	Functional correlation[1]
CAMK2D	2.42	Calcium/Calmodulin Dependent Protein Kinase II Delta	Calcium signaling is crucial for several aspects of plasticity at glutamatergic synapses	ES-I3 cells, primary T cells from blood, adipose cells, fetal muscle trunk
CRBN	2.35	Cereblon	Encodes protein that is found in the cytoplasm localized with a calcium channel membrane protein, and is thought to play a role in brain development	Neuronal progenitor cells; mesoderm, ectoderm, and endoderm cultured cells

1. This information was gathered from Broad Institute and MIT 2015; Weizmann Institute of Science 2019.
2. N/A represents no functional correlation.

Table 3. Genes with highest signal of natural selection for Mexican Maya

Gene	PBS Score	Description[1]	Phenotype[1]	Functional correlation[1]
GTSCR1	2.05	Gilles De La Tourette Syndrome Chromosome Region	Tourette's Disease associated with this gene	Umbilical and placental cells
MIR548AG1	2.05	microRNA	Regulation of post-transcriptional gene regulation	Adipose cells
SLC6A15	1.99	Solute Carrier Family 6 Member 15	Involved in the following pathways: pathways that are transport of glucose, bile salts, organic acids, and metal ions; associated with major depression	Adipocyte cells
LINC01370	1.98	Long Intergenic Non-Protein Coding RNA 137	Associated with diabetes mellitus	Pancreatic islet cells
ZNF385B	1.80	Zinc finger protein 385B	Variants in this gene associated with diabetes, cholesterol, lipoproteins, very low density lipids (VLDL), subcutaneous fat, and Parkinson disease	N/A[2]
VENTXP7	1.78	VENT homeobox pseudogene 7	Encodes genes involved with early embryonic development	N/A
NRG1	1.77	Neuregulin 1	Encodes glycoprotein that mediates cell-cell signaling; dysregulation of this gene associated	N/A

Gene	PBS Score	Description[1]	Phenotype[1]	Functional correlation[1]
			with cancer, schizophrenia, and bipolar disorder	
IZUMO3	1.74	Izumo Sperm-Egg Fusion Protein 3	Involved in fertilization	Endoderm cultured cells
CYP2C9	1.74	Cytochrome P450 Family 2 Subfamily C Member 9	Involved in metabolism; synthesis of cholesterol, steroids and other lipids	N/A
CPNE3	1.73	Copine 3	Involved in metabolism, immune system pathways	Fetal lung cells, primary T cells from blood, bone marrow, adipose, Ganglion Eminence-derived and cortex derived primary cultured neurospheres, thymus, brain cells, skeletal muscle, colon and rectal smooth muscle cells, colonic mucosa, fetal kidney cells, pancreas, lung, spleen, umbilical cells
ART1	1.72	ADP-Ribosyltransferase 1	Involved in immune system pathways	HUES6 Cells
SOD2	1.71	Superoxide Dismutase 2	Encodes protein that binds toxic products of oxidative phosphorylation	Fetal lung cells, ES-I3 cells, neuron cultured cells, Primary T cells from blood, bone marrow, adipose tissue, brain hippocampus middle, brain substantia nigra, brain anterior caudate, brain cingulate gyrus, adipose, fetal muscle cells, intestine cells, rectal

Gene	PBS Score	Description[1]	Phenotype[1]	Functional correlation[1]
				mucosa, placenta, pancreas, liver, monocytes
ADD3	1.66	Adducin 3	Regulation of blood pressure; related to cAMP-dependent pKA activation, transport of glucose, bile salt, and metal ion	Fetal lung cells, primary T cells from blood, bone marrow, brain tissue cells, thymus, fetal thymus, skeletal and muscle cells, heart cells, intestine cells, fetal kidney and lung, placental cells, umbilical cells
OTOP1	1.64	Transmembrane protein belonging to otopetrin domain protein family	Required for the formation of otoconia and otoliths, calcium carbonate biominerals within the inner ear of mammals that are required for the detection of linear acceleration and gravity; defect in gene associated with balance defect; component of a counterinflammatory pathway that attenuates obesity-induced adipose tissue inflammation and plays an adaptive role in maintaining metabolic homeostasis in obesity	N/A
UGT2B15	1.24	UDP Glucuronosyltransferase Family 2	Encodes a glycosyltransferase that is involved in the metabolism and elimination of toxic compounds; regulation of estrogens and androgens	N/A

1. This information was gathered from Broad Institute and MIT 2015; Weizmann Institute of Science 2019.
2. N/A represents no functional correlation.

Table 4. Genes with highest signal of natural selection for Aymara

A subset of the results of the selection scan done with Mexican Maya as the target population is outlined above (Table 3). The top one hundred genes were analyzed, and those with relevant phenotype and functional correlations were illustrated in Table 3. Several genes were associated with calcium regulation, brain development, olfactory receptors, and the nervous system (RNU5F-1, CDH18, OR52Z1, ROBO2, RAPGEF2, CNTN4, CAMK2D, CRBN). Further genes were associated with the eye (ASAH2B). Lastly, some genes involved with metabolism and reproduction (CCDC73, WDR20, KRTAP9-9) were connected to high PBS scores. The functional correlations for certain genes are connected to a multitude of functions, including but not limited to fetal adrenal gland and muscle trunk cells; adipose cells; muscle and skeletal tissue; primary T cells in blood; brain cells; neuronal progenitor cells; and mesoderm, ectoderm, and endoderm cultured cells. Genes correlated to these functions are CRBN, CAMK2D, CNTN4, ROBO2, WDR20, and RNU5F-1.

As a control, a selection scan was done for the Aymara as well. Previous data analyzing natural selection in the Aymara to high altitude environments has been done (Lindo et al. 2018). If the results of this control support previous literature, then it lends validity to the natural selection scans done in this thesis. The results are outlined in Table 4. The genes with highest PBS scores for the Aymara were associated with metabolism, early development, and reproduction. Notably, in terms of metabolism and the immune system, variants of multiple genes (UGT2B15, OTOP1, ZNF385B, ADD3, CPNE3, SLC6A15, LINC01370) were connected with diabetes; cholesterol regulation; transport of glucose; bile salts, and metal ions; and attenuation of adipose tissue inflammation. Second, gene IZUMO3 was associated with spermatogenesis, as well as sperm-egg fusion. The functional correlations for certain genes are connected to a multitude of functions, including but not limited to fetal lung cells; adipocytes;

muscle and skeletal tissue; primary T cells in blood; brain cells; digestive tract cells; and umbilical and placental cells. Genes correlated to these functions are GTSCR1, MIR548AG1, SLC6A15, LINC01370, CPNE3, SOD2, and ADD3.

V. Discussion

As the purpose of this study is to formulate a population history for the Maya of Mexico, the results of the analysis must be compiled into an overarching statement describing this history. The purpose of the PCA was to determine inter-relatedness between the Mexican Maya of interest and other populations in the Americas. Populations that cluster nearer one another in the PCA allude to greater similarity between populations, and thus, historically more mating and interaction. The results of the PCA follow previous research done in the field in that North, Central, and South America are similar within group, and they follow a spectrum across groups (Bolnick et al. 2016; Bryc et al. 2010; Raghavan et al. 2015). This corroborates previous research that states that as humans came into the Americas via the Bering Strait, they quickly populated all of the Americas, differentiating into distinct groups at around 13,000 years ago (Raghavan et al. 2015). The PCA shows these historical and geographical trends. In terms of specifically the Maya, they too group within the Mesoamerican groups. The other indigenous groups in the same Mexican region also follow this trend, showing that despite modern cultural differences, as well as some genetic differences, generally indigenous groups of Mexico are more similar to each other than to other regions of the Americas.

Some groups break from this general trend, which in some cases also corroborates previous research done in this region. Both modern and ancient Brazilian groups (i.e. Surui, Sumidouro) group further from this main spectrum of North, Central, and South American genomes on the PCA. While controversial due to the data being mitochondrial DNA, past

research has seen evidence of Polynesian or Oceanian DNA in Brazilian genomes (Malaspinas et al. 2014). This discrepancy in Brazilians versus other groups of the Americas could very likely be prescribed to this phenomenon. Furthermore, a gap exists between ancient and modern Brazilians. Multiple theories could explain this. First, there could have been a large population replacement between the time period of the Sumidouro population and the modern Surui. Second, as Brazil is home to many diverse indigenous groups, the ancient Sumidouro may not have been directly ancestral to the Surui of today. Future research directions could explore these questions, but as the focus of this study is Mesoamerican populations, this discussion is beyond the scope of this thesis.

Another interesting connection is the similarity of ancient and modern Andeans (including modern Quechua and Aymara) to some populations at the focus of this study. Notably, the Tarahumara of Mexico group more closely to both Andeans and North Americans, such as Piapoco and Upper Sun River Valley (USR). These trends could be explained by geography in some sense. The Tarahumara are a group in northern Mexico, close to the United States border. The Piapoco are found in the southwestern United States. Therefore, geography supports this relationship because if these groups are nearer to one another, they are more likely to interact and produce demographic changes. However, not only do the Tarahumara group closely to North Americans, but they also group with Andeans, specifically the ancient Andeans and Aymara. Again, geography and migratory patterns could explain this to some extent. Groups in this study labeled “Andeans” were done so because of their proximity to the Andes Mountains. This mountain range thus would cause great isolation to these groups. As the widely expected theory is that as humans entered the Americas, they descended further south, the groups that entered the Andes region may have thus become isolated (Harris et al. 2017). Migration

between groups could have been less so than their Mesoamerican and South American counterparts. For instance, previous archaeological, linguistic, and genetic research has shown the active migratory patterns found in Mayan and other indigenous populations in Mesoamerica (Little 2008; McKillop 2004). This claim is further corroborated by the PCA plot, as Mesoamericans sit between northern and southern populations. However, Andeans may have experienced less migration, and thus, less differentiation. This could explain the similarity to groups further north, such as the Piapoco, the Tarahumara, and USR. Thus, in terms of possible migration and mating patterns of the Maya and indigenous populations of interest to this study, the results of the PCA do support the hypothesis that the Maya and other groups interacted with one another enough to have substantial genomic similarity, yet retained some differences as distinct cultural groups.

Looking at the natural selection scan done for both the Aymara and the Mexican Maya, fascinating conclusions can be made in regard to adaptations to their environments. In terms of the Aymara, these results support past research that has shown adaptation to the low oxygen climate of the Andes (Lindo et al. 2018; Harris et al. 2017). For example, the Aymara, as opposed to the Maya whom live in a lower altitude, have high selection for genes which control metabolism (for example, cholesterol and fat regulation). This could be an adaption to the cold climate of the Andes (Lindo et al. 2018; Harris et al. 2017). Lastly, several genes were also selected for which controlled fertilization and early embryonic development. The selection for these genes would support the idea that adaptions in fertility would prove beneficial to a population's long-term survival in their specific environment. This is supported by functional correlation. Many genes are connected to fetal lung tissue, adipose cells, blood cells, umbilical cells and brain cells. These can all be tied to the transport of oxygen through the body (via the

lungs, blood, and brain), as well as the retention of fat. This conclusion is further supported by the cold, high altitude climate in which the Aymara reside today.

In terms of the selection scan done on the Mexican Maya, an interesting trend appears. Multiple genes were selected that encode for controlling calcium signaling, brain development, and axon growth. Calcium is a critical ion in cell-cell signaling in the human body. Furthermore, its function has been tied to autism spectrum disorders (Breitenkamp et al. 2015; Piton et al. 2008). Notably, one gene in this population that was highly selected is also tied to autism, as well as proteins expressed in the central nervous system, the eye, and olfactory and sensory proteins. These genes are also tied to relevant biological functional correlation in human tissue, including brain cells and neuronal progenitor cells, as well as adrenal cells. Thus, it is likely that there has been some selection on neural development and hormonal signaling in this population.

To explain why selection might have been working on neural development and hormonal signaling in the Mayan sample, there is some correlation between the Mexican Maya and an adaption to the function of calcium in the body. A possible explanation of this trend could lie in the nixtamalizaton of foods in Mesoamerica, a process that dates back to around 4,000 years ago (Staller et al. 2010). Nixtamalization greatly increases the calcium and niacin content of maize, the staple crop of Mesoamerica, which otherwise can be nutritionally poor (Wacher 2003). The adaption to calcium binding and regulation could be an effect of two potential situations. First, pre-nixtamaliztion, a diet low in calcium and other nutrients could have caused an adaption for increased calcium binding and regulation. Alternatively, post-nixtamalization, a diet with increased calcium could have caused an adaption for decreased calcium binding and regulation.

The strengths of this study are that it is a comprehensive comparison of both modern and ancient Native American genomes to the Mexican Maya population. Furthermore, it sheds more

light on an underrepresented area of research in population genetics, the Americas. It also provides a cohesive connection between genomics analysis and past archaeological and historical data. Limitations of this study are that there was a smaller sample size of ancient Native American genomes, simply due to availability. Along the same line of thought, more power could be added to the conclusions of this study with a greater sample size of Mayan populations, from beyond Mexico. Again, because of a lack of available genomes, these additional data are not unavailable.

VI. Conclusion

The goal of this study was to provide more research on an underrepresented region of the world, as well as to provide a picture of Mayan population history and adaption. Through the use of principal component analysis (PCA) and the population branch statistic (PBS), I am able to show that Maya interacted with groups near them ranging from North to South America, leading to demographic changes reflected in the modern genome. The Maya also potentially adapted to their diet, as is shown in the natural selection scan results.

As the PCA illustrated, the Maya of Mexico interacted with groups in both northern and southern America, creating a spectrum of genetic similarity that correlates to geography. As is supported by archaeology and history, these mating patterns were driven by trade, war, and ecological climate. Coupled to these factors, the Maya adapted to their environment, as is shown by the results of the selection scan, which point to an adaption to calcium. These results answer some questions regarding modern Mayan population history, yet they also add a new dimension of information, providing a foundation for future studies in Mayan adaption to their environment.

To further this line of inquiry, additional research should be done in the collection and sequencing of both ancient and modern Native American genomes. By increasing the number of

genomes available for analysis, a more accurate representation can be made of the population history of a region. Furthermore, another line of inquiry could evaluate the natural selection of other maize-dependent populations in other regions of the globe to ascertain if the trends found in this study do correlate with diet.

VI. References

Adoga, Moses P, Segun A Fatumo, and Simon M Agwale. 2014. H3Africa: a tipping point for a revolution in bioinformatics, genomics and health research in Africa. *Source Code for Biology and Medicine* 9 (10).

Asturias, Miguel Angel. 1975. *Men of Maize*: Dell Publishing Co, Inc.

Asturias, Miguel Angel. 1989. *Hombres de Maíz*: Editorial Universitaria Centroamericana.

Batz, Giovanni. 2014. Maya Cultural Resistance in Los Angeles: The Recovery of Identity and Culture among Maya Youth. *Latin American Perspectives* 41 (3):194-207.

Bolnick, Deborah A, Jennifer A Raff, Lauren C Springs, Austin W Reynolds, and Aida T Miró-Herrans. 2016. Native American Genomics and Population Histories. *Annual Review of Anthropology* 45:319-340.

Broad Institute, MIT. 2015. HaploReg 4.1. https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php

Bryc, Katarzyna, Christopher Velez, Tatiana Karafet, Andres Moreno-Estrada, Andy Reynolds, Adam Auton, Michael F. Hammer, Carlos D Bustamante, and Harry Ostrer. 2010. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *PNAS* 107 Suppl 2:8954-8961.

Canuto, Marcello A, Francisco Estrada-Belli, Thomas G Garrison, Stephen D Houston, Mary Jane Acuña, Milan Kováč, Damien Marken, Philippe Nondédéo, Luke Auld-Thomas, Cyril Castanet, David Catelain, Carlos R Chiriboga, Tomáš Drápela, Tibor Lieskovsky, Alexandre Tokovinine, Antolín Velasquez, Juan C Fernández-Díaz, and Ramesh Shrestha. 2018. Ancient lowland Maya complexity as revealed by airborne laser scanning of northern Guatemala. *Science* 361 (1355).

Castaneda, Quetzil. 2004. "We Are Not Indigenous!": An Introduction to the Maya Identity of Yucatan. *Journal of Latin American Anthropology* 9 (1):36-63.

Daly, Bobby, and Olufunmilayo Olopade. 2015. A perfect storm: How tumor biology, genomics, and health care delivery patterns collide to create a racial survival disparity in breast cancer and proposed interventions for change. *CA: A Cancer Journal for Clinicians* 65:221-238.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, Gilean McVean, Richard Durbin, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156-2158.

Dat Tran, Tat, Julian Hofrichter, and Jürgen Jost. 2012. An introduction to the mathematical structure of the Wright–Fisher model of population genetics. *Theory in Biosciences* 132:73-82.

Excoffier, Laurent, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C Sousa, and Matthieu Foll. 2013. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics* 9 (10).

Fischer, Edward F. 1999. Cultural Logic and Maya Identity: Rethinking Constructivism and Essentialism. *Current Anthropology* 40 (4):473-500.

Ford, Anabel, Keith C Clarke, and Gary Raines. 2009. Modeling Settlement Patterns of the Late Classic Maya Civilization with Bayesian Methods and Geographic Information Systems. *Annals of the Association of American Geographers* 99 (3):496-520.

Gonzalez-Martin, Antonio, Amaya Gorostiza, Lucia Regalado-Liu, Sergio Arroyo-Pena, Sergio Tirado, Ismael Nuno-Arana, Rodrigo Rubi-Castellanos, Karla Sandoval, Michael D Coble, and Hector Rangel-Villalobos. 2015. Demographic History of Indigenous Populations in Mesoamerica Based on mtDNA Sequence Data. *PloS One* 10 (8).

Harris, Daniel N, Wei Song, Amol C Shetty, Kelly S Levano, Omar Caceres, Carlos Padilla, Victor Borda, David Tarazona, Omar Trujillo, Cesar Sanchez, Michael D Kessler, Marco Galarza, Silvia Capristano, Harrison Montejo, Pedro O Flores-Villanueva, Eduardo Tarazona-Santos, Timothy D O'Connor, and Heinner Gui. 2017. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *PNAS* 115 (28).

Ibarra-Rivera, Lisa, Sheyla Mirabal, Manuela M Regueiro, and Rene J Herrera. 2008. Delineating Genetic Relationships Among the Maya. *American Journal of Physical Anthropology* 135:329-347.

Kistler, S Ashley. 2014. Murder, Memory, and the Maya. *Latin American Research Review* 49 (1):251-260.

Leventhal, Richard M, Carlos Chan Espinosa, and Cristina Coc. 2012. The Modern Maya and Recent History. *Expedition* 54 (1).

Li, Heng. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21): 2987-2993.

Li, Heng, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, 1000 Genomes Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-9.

Lindo, J., Haas, R., Hofman, C., Apata, M., Moraga, M., Verdugo, R. A., et al. (2018).

The genetic prehistory of the Andean highlands 7000 years BP through European contact. *Science Advances*, *4*(11), eaau4921. http://doi.org/10.1126/sciadv.aau4921

Little, Walter E. 2008. A Visual Political Economy of Maya Representations in Guatemala, 1931–1944. *Ethnohistory* 55 (4).

Malaspinas, Anna-Sapfo, Oscar Lao, Hannes Schroeder, Morten Rasmussen, Maanasa Raghavan, Ida Moltke, Paula F Campos, Francisca Santana Sagredo, Simon Rasmussen, Vanessa F Gonçalves, Anders Albrechtsen, Morten E Allentoft, Philip LF Johnson, Mingkun Li, Silvia Reis, Danilo V Bernardo, Michael DeGiorgio, Ana T Duggan, Murilo Bastos, Yong Wang, Jesper Stenderup, J Victor Moreno-Mayar, Soren Brunak, Thomas Sicheritz-Ponten, Emily Hodges, Gregory J Hannon, Ludovic Orlando, T Douglas Price, Jeffrey D Jensen, Rasmus Nielsen, Jan Heinemeier, Jesper Olsen, Claudia Rodrigues-Carvalho, Marta Mirazon Lahr, Walter A Neves, Manfred Kayser, Thomas Higham, Mark Stoneking, Sergio DJ Pena, Eske Willerslev. 2014. Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil. *Current Biology* 24(21): R1035-R1037.

McKillop, Heather. 2004. *The Ancient Maya: New Perspectives*. Santa Barbara, California: ABC CLIO.

Medina, Laurie Kroshus. 2003. Commoditizing culture: Tourism and Maya Identity. *Annals of Tourism Research* 30 (2):353-368.

Nielsen, Rasmus, and Montgomery Slatkin. 2013. *An Introduction to Population Genetics: Theory and Applications*. Sunderland, MA: Sinauer Associates, Inc.

Ochoa-Lugo, Mirna Isabel, María de Lourdes Muñoz, Gerardo Pérez-Ramírez, Kristine G Beaty, Mauro López-Armenta, Javiera Cervini-Silva, Miguel Moreno-Galeana, Adrián Martínez Meza, Eduardo Ramos, Michael H. Crawford, and Arturo Romano-Pacheco. 2016. Genetic Affiliation of Pre-Hispanic and Contemporary Mayas through Maternal Linage. *Human Biology* 88 (2):136-167.

Oland, Maxine, and Joel Palka. 2016. The perduring Maya: new archaeology on early Colonial transitions. *Antiquity* 90 (350):472-486.

Orlando, Ludovic, and Alan Cooper. 2014. Using Ancient DNA to Understand Evolutionary and Ecological Processes. *Annual Review of Ecology, Evolution, and Systematics* 45:573-598.

Palotie, A, and S Ripatti. 2017. Finland establishing the internet of genomics and health data. *Duodecim* 133 (8).

Posth, Cosimo, Nathan Nakatsuka, Iosif Lazaridis, Pontus Skoglund, Swapan Mallick, Thiseas C Lamnidis, Nadin Rohland, Kathrin Nagele, Nicole Adamski, Emilie Bertolini, Nasreen Broomandkhoshbacht, Alan Cooper, Brendan J Culleton, Tiago Ferras, Matthew Ferry, Anja Furtwangler, Wolfgang Haak, Kelly Harkins, Thomas K Harper, Tabita Hunemeier,

Anne Marie Lawson, Bastien Llamas, Megan Michel, Elizabeth Nelson, Jonas Oppenheimer, Nick Patterson, Stephan Schiffels, Jakob Sedig, Kristin Stewardson, Sahra Talamo, Chuan-Chao Wang, Jean-Jacques Hublin, Mark Hubbe, Katerina Harvati, Amalia Nuevo Delaunay, Judith Beier, Michael Francken, Peter Kaulicke, Hugo Reyes-Centeno, Kurt Rademaker, Willa R Trask, Mark Robinson, Said M Gutierrez, Keith M Prufer, Domingo C Salazar-Garcia, Eliane N Chim, Lisiane Muller Plumm Gomes, Marcony L Alves, Andersen Liryo, Mariana Inglez, Rodrigo E Oliveira, Danilo V Bernardo, Alberto Barioni, Veronica Wesolowski, Nahuel A Scheifler, Mario A Rivera, Claudia R Plens, Pablo G Messineo, Levy Figuti, Daniel Corach, Clara Scabuzzo, Sabine Eggers, Paulo DeBlasis, Markus Reindel, Cesar Mendez, Gustavo Politis, Elsa Tomasto-Cagigao, Douglas J Kennett, Andre Strauss, Lars Fehren-Schmitz, Johannes Krauase, and David Reich. 2018. Reconstructing the Deep Population History of Central and South America. *Cell* 175:1185-1197.

Raghavan, Maanasa, Matthias Steinrucken, Kelley Harris, Stephan Schiffels, Simon Rasmussen, Michael DeGiorgio, Anders Albrechtsen, Cristina Valdiosera, Maria C Avila-Arcos, Anna-Sapfo Malaspinas, Anders Eriksson, Ida Moltke, Mait Metspalu, Julian R Homburger, Jeff Wall, Omar E Cornejo, J Victor Moreno-Mayar, Thorfinn S Korneliussen, Tracey L. Pierre, Morten Rasmussen, Paula F. Campos, Peter de Barros Damgaard, Morten E Allentoft, John Lindo, Ene Metspalu, Ricardo Rodriguez-Varela, Josefina Mansilla, Celeste Henrickson, Andaine Seguin-Orlando, Helena Malmstrom, Thomas Stafford Jr, Suyash S Shringarpure, Andres Moreno-Estrada, Monika Karmin, Kristiina Tambets, Anders Bergstrom, Yali Xue, Vera Warmuth, Andrew D Friend, Joy Singarayer, Paul Valdes, Francois Balloux, Ilan Leboreiro, Jose Luis Vera, Hector Rangel-Villalobos, Davide Pettener, Donata Luiselli, Loren G Davis, Evelyne Heyer, Christoph PE Zollikofer, Marcia S Ponce de Leon, Colin I Smith, Vaughan Grimes, Kelly-Anne Pike, Michael Deal, Benjamin T Fuller, Bernando Arriaza, Vivien Standen, Maria F Luz, Francois Ricaut, Niede Guidon, Ludmila P. Osipova, Mikhail I Voevoda, Olga L Posukh, Oleg Balanovsky, Maria Lavryashina, Yuri Bogunov, Elza Khusnutdinova, Marina Gubina, Elena Balanovska, Sardana A. Fedorova, Sergey Litvinov, Boris Malyarchuk, Miroslava Derenko, MJ Mosher, David Archer, Jerome C. Cybulski, Barbara Petzelt, Joycelynn Mitchell, Rosita Worl, Paul J Norman, Peter Parham, Brian M Kemp, Toomas Kivisild, Chris Tyler-Smith, Manjinder S Sandhu, Michael H. Crawford, Richard Villlems, David Glenn Smith, Michael R Waters, Ted Goebel, John R Johnson, Ripan S. Malhi, Mattias Jakobsson, David J Meltzer, Andrea Manica, Richard Durbin, Carlos D Bustamante, Yun S Song, Rasmus Nielsen, and Eske Willerslev. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* 349 (6250).

Sánchez-Pozos, Katy, María Guadalupe Ortíz-López, Bárbara I Peña-Espinoza, María de los Angeles Granados-Silvestre, Veronica Jimenez-Jacinto, Jerome Verleyen, Fasil Tekola-Ayele, Alejandro Sanchez-Flores, and Marta Menjivar. 2018. Whole-exome sequencing in Maya indigenous families: variant in PPP1R3A is associated with type 2 diabetes. *Molecular Genetics and Genomics* 293:1205-1216.

Schuster, Angela MH. 1997. Ristuals of the Modern Maya. *Archaeology* 50 (4):50-53.

Shapiro, Beth, and Michael Hofreiter. 2014. A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. *Science* 343 (6169).

Slatkin, Montgomery, and Fernando Racimo. 2016. Ancient DNA and human history. *PNAS* 113 (23).

Söchtig, Jens, Vanesa Álvarez-Iglesias, Ana Mosquera-Miguel, Miguel Gelabert-Besada, Alberto Gómez-Carballa, and Antonio Salas. 2015. Genomic insights on the ethno-history of the Maya and the 'Ladinos' from Guatemala. *BMC Genomics* 16 (131).

TallBear, Kim. 2013. Native American DNA. Minneapolis, MN: University of Minnesota Press.

Tataru, Paula, Maria Simonsen, Thomas Bataillon, and Asger Hobolth. 2016. Statistical Inference in the Wright–Fisher Model Using Allele Frequency Data. *Systematic Biology* 66 (1).

Turner, Stephen D. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*.

Wade, Peter. 2017. *Degrees of Mixture, Degrees of Freedom: Genomics, Multiculturalism, and Race in Latin America*. Durham, NC: Duke University Press.

Weizmann Institute of Science. 2019. Genecards. https://www.genecards.org/

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag*.

Yi, Xin, Yu Liang, Emilia Huerta-Sánchez, Xin Jin, Zha Xi Ping Cuo, John E Pool, Xun Xu, Hui Jiang, Nicolas Vinckenbosch, Thorfinn S Korneliussen, Hancheng Zheng, Tao Liu, Weiming He, Kui Li, Ruibang Luo, Xifang Nie, Honglong Wu, Meiru Zhao, Hongzhi Cao, Jing Zou, Ying Shan, Shuzheng Li, Qi Yang, Asan, Peixiang Ni, Geng Tian, Junming Xu, Xiao Liu, Tao Jiang, Renhua Wu, Guangyu Zhou, Meifang Tang, Junjie Qin, Tong Wang, Shuijian Feng, Guohong Li, Huasang, Jiangbai Luosang, Wei Wang, Fang Chen, Yading Wang, Xiaoguang Zheng, Zhuo Li, Zhuoma Bianba, Ge Yang, Xinping Wang, Shuhui Tang, Guoyi Gao, Yong Chen, Zhen Luo, Lamu Gusang, Zheng Cao, Qinghui Zhang, Weihan Ouyang, Xiaoli Ren, Huiqing Liang, Huisong Zheng, Yebo Huang, Jingxiang Li, Lars Bolund, Karsten Kristiansen, Yingrui Li, Yong Zhang, Xiuqing Zhang, Ruiqiang Li, Songgang Li, Huanming Yang, Rasmus Nielsen, Jun Wang, and Jian Wang. 2010. Sequencing of 50 Human Exomes Reveals Adaption to High Altitude. *Science* 329.

Zheng X, Levine D, Shen J, Gogarten S, Laurie C, Weir B. 2012. A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics* 28(24), 3326-3328.