# The Population History of the Caribbean:

# Perspectives from Ancient and Modern DNA Analysis

by

Maria Nieves Colón

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2017 by the
Graduate Supervisory Committee:

Anne Stone, Chair
William Pestle
Jada Benn-Torres
Christopher Stojanowski

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

Although the Caribbean has been continuously inhabited for the last 7,000 years, European contact in the last 500 years dramatically reshaped the cultural and genetic makeup of island populations. Several recent studies have explored the genetic diversity of Caribbean Latinos and have characterized Native American variation present within their genomes. However, the difficulty of obtaining ancient DNA from pre-contact populations and the underrepresentation of non-Latino Caribbean islanders in current research have prevented a complete understanding of genetic variation over time and space in the Caribbean basin. This dissertation uses two approaches to characterize the role of migration and admixture in the demographic history of Caribbean islanders. First, autosomal variants were genotyped in a sample of 55 Afro-Caribbeans from five islands in the Lesser Antilles: Grenada, St. Kitts, St. Lucia, Trinidad, and St. Vincent. These data were used to characterize genetic structure, ancestry and signatures of selection in these populations. The results demonstrate a complex pattern of admixture since European contact, including a strong signature of sex-biased mating and inputs from at least five continental populations to the autosomal ancestry of Afro-Caribbean peoples. Second, ancient mitochondrial and nuclear DNA were obtained from 60 skeletal remains, dated between A.D. 500–1300, from three archaeological sites in Puerto Rico: Paso del Indio, Punta Candelero and Tibes. The ancient data were used to reassesses existing models for the peopling of Puerto Rico and the Caribbean and to examine the extent of genetic continuity between ancient and modern populations. Project findings support a largely South American origin for Ceramic Age Caribbean populations and identify some genetic continuity between pre and post contact islanders. The above study was aided by

development and testing of extraction methods optimized for recovery of ancient DNA from tropical contexts. Overall, project findings characterize how ancient indigenous groups, European colonial regimes, the African Slave Trade and modern labor movements have shaped the genomic diversity of Caribbean islanders. In addition to its anthropological and historical importance, such knowledge is also essential for informing the identification of medically relevant genetic variation in these populations.

# DEDICATION

To the people of the Caribbean.

ACKNOWLEDGMENTS

This dissertation would not have been possible without several people and institutions who provided funding, training, samples and mentorship. I would like to thank the funding sources that financed this research and large part of my graduate education at Arizona State University (ASU). These include the National Science Foundation Graduate Research Fellowship (GRFP), the National Science Foundation Doctoral Dissertation Improvement Grant in Archaeology (DDIG), the Rust Family Foundation Grant for Archeological Research, the Sigma Xi Grants-in-Aid of Research program, the ASU Center for Evolutionary Medicine Venture Fund Grant, the ASU School of International Letters and Cultures Foster Latin American Studies Award, the ASU Graduate and Professional Student Association Graduate Research Support Grant, and the ASU School of Human Evolution and Social Change Dissertation Fellowship and Research Committee Grants.

I would also like to thank my collaborators in this research: Dr. Carlos D. Bustamante for hosting me in his laboratory at Stanford University for training in whole genome enrichment methods and for providing access to computational resources, reagents and sequencing; the postdocs and staff at Stanford University Dr. Meredith Carpenter, Dr. Rosa Fregel, Dr. Maria C. Ávila-Arcos, Dr. Chris Gignoux, Dr. Alicia Martin, Dr. Morten Rasmussen and Alexandra Adams-Sockell for their help and assistance during my time there; Dr. L. Antonio Curet for guidance, mentorship and training in archaeological methods;  Dr. William J. Pestle and Dr. Edwin Crespo-Torres for their help with sampling skeletal remains from Puerto Rico; Dr. Jada Benn-Torres for hosting me in her laboratory and help with sampling Lesser Antillean populations; and

Dr. Andrew Ozga, Kathleen Fowler and Katherine Skerry for their help with labwork related to the chapters 3 and 4 of this dissertation.

I also want to thank all my mentors and people who have helped me along the way: the members of my committee Dr. William J. Pestle, Dr. Jada Benn-Torres, and Dr. Christopher Stojanowski for providing helpful commentary and reading interminable amounts of grant proposal and dissertation chapter drafts; Dr. Juan C. Martinez-Cruzado for provided my first training and mentorship in genetics methods; Dr. Christina Warinner and Dr. Cecil Lewis for hosting me in their lab at University of Oklahoma for training on extraction methods; Dr. Melissa Wilson-Sayres and Dr. Tim Webster for their help with bioinformatics methods; and Dr. Anne C. Stone for being a great adviser, mentor and role model. Thank you as well to ASU support staff who through their work help make all our research possible.

Finally, I want to thank my lab mates, my graduate school colleagues, my family and friends for their enduring support during this journey. And lastly, but perhaps most importantly, thank you to all the people who donated samples for this research. Without their participation, none of this would be possible.

TABLE OF CONTENTS

Page

LIST OF TABLES ........................................ix

LIST OF FIGURES ........................................x

CHAPTER

1 INTRODUCTION ........................................ 1

2 UNDERSTANDING ADMIXTURE, GENETIC STRUCTURE AND ADAPTATION IN AFRO-CARIBBEANS FROM THE ENGLISH-SPEAKING LESSER ANTILLES ........................................6

2.1 Abstract ........................................6

2.2 Introduction ........................................7

2.3 Materials and Methods ........................................15

2.4 Results ........................................30

2.5 Discussion ........................................52

2.6 Conclusion ........................................66

3 COMPARATIVE PERFORMANCE OF ANCIENT DNA EXTRACTION METHODS IN REMAINS FROM TROPICAL ENVIRONMENTS ..... 68

3.1 Abstract ........................................68

3.2 Introduction ........................................69

3.3 Materials and Methods ........................................74

3.4 Results ........................................83

3.5 Discussion ........................................95

3.6 Conclusion ........................................100

Page

CHAPTER

4 RECONSTRUCTING THE PEOPLING OF PRE-CONTACT PUERTO RICO THROUGH ANCIENT DNA ANALYSIS .................................... 102

4.1 Abstract........................................................................................102

4.2 Introduction...................................................................................103

4.3 Materials and Methods .................................................................113

4.4 Results ..........................................................................................130

4.5 Discussion.....................................................................................151

4.6 Conclusion ....................................................................................169

5 CONCLUSION .................. ..................................................................... 172

REFERENCES...... ............................................................................................... 177

APPENDIX

A INSTITUTIONAL REVIEW BOARD APPROVALS AND MATERIAL TRANSFER AGREEMENTS FOR SAMPLES IN CHAPTER 2 ... ... 223

B SUPPLEMENTARY TABLES FOR CHAPTER 2 (SM1 to SM3)................... 224

C SUPPLEMENTARY FIGURES FOR CHAPTER 2 (SM1 TO SM14) ............. 225

D PROTOCOL FOR EXTRACTION METHOD H ........................................... 226

E DAMAGE PLOTS FOR SHOTGUN LIBRARIES IN CHAPTER 3 ............... 232

F DAMAGE PLOTS FOR MTDNA ENRICHED LIBRARIES IN CHAPTER 3 233

G SUPPLEMENTARY TABLES FOR CHAPTER 3 (SM4 TO SM6) ............... 234

H SUPPLEMENTARY FIGURES FOR CHAPTER 3 (SM15 TO SM29) .......... 235

APPENDIX PAGE

I TABLE OF MAPPING STATISTICS AND HAPLOGROUP ASSIGNMENTS FOR 27 ANCIENT SAMPLES ANALYSED IN CHAPTER 4 ..... 236

J SUPPLEMENTARY TABLES FOR CHAPTER 4 (SM7 TO SM22)................. 239

K SUPPLEMENTARY FIGURES FOR CHAPTER 4 (SM30 TO SM34) ............ 240

LIST OF TABLES

Table Page

1. Autosomal Global Ancestry Estimates for Lesser Antilles Afro-caribbean Populations Using RFMix and ADMIXTURE. ..................................................... 34

2. Ancestry Proportions for Chromosome X in Lesser Antilles Afro-caribbean Populations (Average for Females Only). ............................................................ 41

3. Top 1% 100 Kb Candidate Regions Identified in iHS Analysis Conducted with 55 Lesser Antilles Afro-caribbeans. ............................................................... 44

4. Significant Results of PANTHER Statistical Overrepresentation Test on Gene Lists Retrieved from Top 1% Candidate Windows for iHS and XP-EHH Analyses in Lesser Antilles Afro-caribbeans. ............................................................... 48

5. Autosomal Markers with Significant Ancestry Skews. .......................................... 51

6. Ancient and Historical Tooth Samples Included in Chapter 3. .............................. 71

7. Results of Paired Tests to Determine Significant Differences in Sample Means for Shotgun Libraries. ........................................................................................... 84

8. Results of Paired Tests to Determine Significant Differences in Sample Means for mtDNA Enriched Libraries. ........................................................................ 93

9. Diversity Summary Statistics Calculated from Complete mtDNA Data. .............. 134

10. Diversity Summary Statistics Calculated from HVR-1 Data (16024-16391). ...... 135

LIST OF FIGURES

Figure Page

1. Map Of The Caribbean Archipelago with Islands Included in Present Study. .......... 9

2. PCA Performed with 44,219 Autosomal SNPs Intersected Between Lesser Antilles Genotyped Samples and Reference Populations from Africa, Europe Central-South Asia, East Asia, Oceania and the Americas Collected from fhe Literature. PC 1 And 2 Contain 10.32% of Total Variation. ........................................................................ 31

3. PCA Performed with 23,636 Autosomal SNPs Intersected Between Lesser Antilles Genotyped Samples, Reference Populations from Africa, Europe Central-South Asia, East Asia, Oceania and the Americas, and Admixed Caribbean Populations. PC 1 and 2 Contain 8.18% of Total Variation. COL = Colombia, HUR = Honduras, CUB = Cuba, HAI = Haiti, DOM= Dominican Republic, PUR = Puerto Rico, ACB = Afro-caribbean Barbados. ........................................................................ 32

4. ADMIXTURE Clustering Performed with 44,219 Autosomal SNPs Common to Lesser Antilles Genotyped Samples and Reference Populations from the Caribbean, Africa, Europe South Asia, East Asia and the Americas (Merged Dataset 1). See Table SM3 for Abbreviations. ........................................................................ 33

5. ADMIXTURE Clustering Performed with 23,636 Autosomal SNPs Common to Lesser Antilles Genotyped Samples and Reference Populations from the Caribbean, Africa, Europe South Asia, East Asia and the Americas (Merged Dataset 2). See Table SM3 for Abbreviations. ........................................................................ 33

Figure Page

6. Distribution of Continental Ancestry Proportions among Lesser Antilles Afro-caribbeans Generated with ADMIXTURE Analysis At K=5 with 44,219 Autosomal SNPs (Merged Dataset 1)........................................................................... 35

7. Distribution of Continental Ancestry Proportions across 12 Admixed Caribbean Populations Generated from ADMIXTURE Analysis at K=5 with 23,636 Autosomal SNPs (Merged Dataset 2)...................................................................... 36

8. Local Ancestry Karyotypes of Four Lesser Antilles Afro-caribbean Individuals. An Individual With: (A) High Proportions of African Ancestry. (B) High Proportions of South Asian Ancestry. (C) High Proportions of European and East Asian Ancestry. (D) High Proportion of Native American Ancestry. ............................... 39

9. Global Ancestry Proportions Were Averaged per Individual Genome from Rfmix Local Ancestry Per-Marker Estimates. Local Ancestry Estimates Were Conducted with Merged Dataset 5 (119,277 Autosomal SNPs)............................................... 40

10. (A) Boxplot Comparison of Gobal Ancestry Proportions Derived from Local Ancestry Estimation. Significant Comparisons in Wilcoxon Paired Test Noted with Asterisks: *** = $P \leq 0.001$, ** = $P \leq 0.01$, * = $P \leq 0.05$, ns = $P > 0.05$. (B) Stacked Bar Plot of Autosomal Global Ancestry Proportions. (C) Stacked Bar Plot of Chromosome X Ancestry Proportions.................................................................... 42

Figure Page


11. Manhattan Plot of Standardized Genome-Wide |iHS| Scores for Lesser Antilles Afro-caribbeans (N=55). The Blue Line Indicates the Theoretical Threshold of |iHS| > 2 and the Red Line Indicates the Empirically Determined Top 1% Quantile Threshold. SNPs in Green Are Within Top 1% 100 Kb Candidate Windows on Chromosomes 1-3,6,8, 17 and 19. .......... 45
12. Manhattan Plots of Standardized Genome-wide XP-EHH Scores Across Four Cross-Population Comparisons. Values above the Red Line Were Empirically Determined to Be in the Top 1% Quantiles. .......... 46
13. Genome-wide Local Ancestry Proportions per Chromosome in Lesser Antilles Afro-caribbeans (N=55). African Ancestry is >50% for All Loci (119,277 Autosomal SNPs). .......... 50
14. Standardized Genome-wide Local Ancestry Deviations in Lesser Antilles Afro-caribbeans (N=55). The Dotted Lines Represent 4 Standard Deviations from Average Global Ancestry Proportions (119,277 Autosomal SNPs). .......... 51
15. Schematic of Extraction Methods Tested in Chapter 3. .......... 72
16. Sampling Locations for Ancient and Historic Tooth Remains Studied in Chapter 3. .......... 75

Figure Page


17. Per Sample Comparison of Raw DNA Yields after Extraction and DNA Content in Shotgun Libraries. (A) Double Bar Graph Comparing Extraction Yields (Measured as ng/uL on Qubit High Sensitivity Assay) for Samples Extracted with Both Methods. (B) Double Bar Graph Comparing Mean Cycle Threshold Values (Ct) Measuring DNA Content in Shotgun Libraries. .................................................... 84
18. Per Sample Endogenous Content in Shotgun Libraries. Percent Endogenous Content Calculated as Number of Mapped, Unique Reads Divided over Total Down Sampled Reads. ........................................................................................ 86
19. Scatterplots Depicting Relationship Between Endogenous Content, Clonality and Complexity in Shotgun Libraries. (A) Clonality or Duplicate Rate as Number of Duplicate Reads Divided over Total Downsampled Reads. (B) Complexity as the Number of Distinct Reads Before Duplicate Removal (Measured Using Preseq). . 87
20. Extrapolation Curves for Shotguns Library Complexity Estimation. Curves Are Shown for the Two Samples with Highest Number of Reads: PI-67 and GB-7. Top Inset Shows Zoomed in Results for PI-67. Extrapolation Curve and Confidence Interval Estimation Was Performed in Preseq Using a Step Size of 100,000 and Default Sequencing Effort of 10 Billion Reads. The Dotted Line Denotes the Number of Reads Randomly Downsampled for Each Sample Pair: 5.1 Million Reads for GB-7 and 6.8 Million Reads for PI-67. .................................................. 88
21. DNA Fragment Length Distributions for Unique, Mapped Reads in Shotgun Libraries. ......................................................................................................... 90

Figure Page

22. Boxplots Comparing Distributions of DNA Fragment Length (A) and %GC Content (B) Values in Shotgun Libraries. ........................................................................ 90

23. Per Sample Comparison of mtDNA Enrichment. Enrichment Factor Shown in Log10 for Ease of Graph Interpretation. After Downsampling, Replicate PI-388 Extracted with Method H had Zero Reads Mapping to the Mitochondrial Genome. ........................................................................................................................ 92

24. Distribution of Average DNA Fragment Length (A) and Average %GC Content Values (B) for Mtdna Enriched Libraries. ............................................................. 94

25. Map of Puerto Rico and the Caribbean Islands. Islands for Which Pre-contact Ancient Mitochondrial DNA Data Are Available Are Labeled in Blue. Inset Zooms into Puerto Rico to Demonstrate Approximate Location of Three Archaeological Sites from Which Skeletal Remains Were Sampled. Sample Size for Each Site is Noted. Total Sample Size is N=60 (Images Modified from WorldAtlas). ............ 104

26. Simplified timeline of the pre-contact history of the Antilles demonstrating occupation periods of studied sites. Site chronologies based on published sources. Dashed lines represent the non-continuous occupation of Paso del Indio. Red boxes encompass minimum and maximum median calibrated radiocarbon dates (cal A.D.) for sampled skeletal remains. ..................................................................... 115

27. Per Site Comparison of Recovery Rates for Complete mtDNA Data after Enrichment. Samples Were Included in Subsequent Analyses When Read Depth >3X and Genome Coverage >93%. MT = mtDNA. ............................................ 132

Figure Page

28. *Fst* Genetic Distances from Complete mtDNA Sequence Data. (A) MDS Plot in Two Dimensions. Surui and Karitiana Were Excluded from the Plot Due to Falling Extremely Far from Other Populations. (B) Heatmap Matrix of Raw Pairwise *Fst* Values. Darker Colors Represent Increased Genetic Differentiation. PC = Pre-contact. .......... 137

29. *Fst* Genetic Distances from HVR-I Sequence Data from Ancient (A) and Modern (B) Caribbean Populations. Heatmap Matrix of Raw Pairwise *Fst* Values. Darker Colors Represent Increased Genetic Differentiation. PC = Pre-contact. .......... 138

30. Distribution of mtDNA Sub-haplogroups over Time Across All Archaeological Sites. Temporal Groups Identified Through Hierarchical Clustering of Median calAD Radiocarbon Dates for 21Samples. Clusters are: (1) 491-671 Median calAD with n=6 Individuals, (2) 824-1039 calAD with n=9, and (3) 1095-1270 calAD with n=6. .......... 140

31. Complete mtDNA Networks for Haplogroups A2 and C1 Including Pre and Post Contact Puerto Rico Populations. Mutations Identified Relative to the rCRS. Node Size is Proportional to the Frequency of a given Haplotype. Yellow Arrows Indicate Lineages C1b2 and A2+16218 Which Were Included in Coalescent Dating Analysis. PC = Pre-contact. MV = Median Vector .......... 142

Figure Page


32. HVR-I Networks for Haplogroups A2, C1 and D1 Including Pre and Post Contact Caribbean Populations. Networks Are Restricted to Positions 16056 – 16391 Which Were Sequenced Across All Datasets. Mutations Identified Relative to the rCRS. Node Size Is Proportional to the Frequency of a Given Haplotype. PC = Pre-contact ............ 145
33. Damage Plots and Fragment Misincorporation Patterns for Autosomal Reads Recovered from Sample PI420 after WISC Enrichment. Plot Generated with mapDamage. ............ 148
34. PCA Conducted with 11,742 Autosomal SNPs Comparing PI420 to 51 HGDP Populations (PC 1 vs 2). Sample PI420 Shown as a Yellow Star............ 149
35. ADMIXTURE Analysis K=2 to K=6 with 11,090 Autosomal SNPs Comparing PI420 to HGDP Populations. SNPs Were Thinned for Linkage Disequilibrium. PI420 Sample is the Last Bar on the Right. ............ 150

# CHAPTER 1: INTRODUCTION

This dissertation examines the role of migration and genetic admixture in the history and evolution of human populations from the Caribbean islands. In evolutionary genetics, migration or gene flow is defined as movement between groups that leads to genetic exchange (Hedrick 2011, Slatkin 1985). As one of the four forces of evolution, gene flow can have a large effect on the origin and demography of populations. It can introduce new genetic variation, homogenize previously isolated groups and disrupt or enable the effects of other forces such as genetic drift and natural selection (Slatkin 1985, Cavalli-Sforza and Feldman 2003). For instance, migration and subsequent admixture were major drivers of change and adaptation for the human lineage (Wells and Stock 2012). As they left Africa, modern humans interbred with other (now extinct) hominin species and through this process obtained advantageous genetic variants (Huerta-Sanchez et al. 2014, Deschamps et al. 2016, Dannemann, Andres, and Kelso 2016, Sankararaman et al. 2014, Reich et al. 2011). Exchange of locally adaptive variants also occurred in the more recent past between modern human groups. Examples include introgression of alleles involved in high-altitude adaptation from ancestral Sherpa and Chinese populations into Tibetans (Jeong et al. 2014) and introgression of adaptive skin pigmentation alleles from Bantu agriculturalists into Khoe-San hunter gatherers (Schlebusch et al. 2012). At least 1,000 instances of genomic admixture have been identified in the demographic history of our species during the last 4,000 years (Hellenthal et al. 2014). Therefore, characterizing how gene flow shapes our genetic diversity is vital for understanding human evolution, history and health and is a major focus of modern genomics research (O'Rourke 2012).

Until recently, most research into the role of migration in human population history focused on analysis of uniparental markers (such as the Y-chromosome or mitochondrial DNA) or small fragments of the autosomal genome. But in the last two decades, improvements in sequence quality and the decreasing costs afforded by next-generation sequencing technologies have led to the incorporation of complete genomes into anthropological and evolutionary genetics research (Mardis 2008, Pugach and Stoneking 2015, Metzker 2010). This growth has also included the improved ability to recover ancient DNA and the beginning of the field of paleogenomics. The data recovered from ancient sources have challenged previous understanding of interaction, migration and admixture among ancient human groups and their relationships to extant populations (Pickrell and Reich 2014, Slatkin and Racimo 2016). For example, Raghavan et al. (2015) and Skoglund et al. (2015) proposed two alternative hypotheses for the peopling of the Americas. Both hypotheses seek to explain the recent discovery of a shared genomic component between Amazonian and Australo-Melanesian populations. Neither insight would have been possible without the consideration of both ancient and modern genomics data.

In this dissertation, I implement a similarly combined approach using both ancient and modern genomic data to explore signatures of migration, admixture and interaction among ancient and extant Caribbean populations. Approximately seven thousand years ago, the Caribbean became the last region of the Americas to be settled by humans (Pantel 2003, Reid 2014). But, native island populations were the first American groups to encounter European colonizers, experience the disease transfer of the Columbian exchange and interact with the victims of the Atlantic Slave trade (Cook 1998, Sheridan

1972, Whitehead 1999). Thus, due to this complex history of extensive and continuous exchange, Caribbean islanders present an excellent case for studying how admixture between distantly related parental populations can shape neutral and adaptive genetic variation.

However, a complete understanding of human genomic diversity across the Caribbean region is hampered by sampling gaps of both past and present populations. Due to the difficulties of obtaining ancient DNA from the tropics, the genetic diversity of pre-contact Caribbean groups is not well characterized. Efforts have been made to address this problem by studying Native American fragments in the genomes of admixed islanders (Bryc, Velez, et al. 2010, Gravel et al. 2013, Moreno-Estrada et al. 2013, Martínez-Cruzado 2002, 2010, Martínez-Cruzado et al. 2005, Via et al. 2011, Mendizabal et al. 2008, Marcheco-Teruel et al. 2014, Benn-Torres, Kittles, and Stone 2007, Benn-Torres et al. 2015, Vilar et al. 2014). But, because modern populations do not retain all the genomic diversity of ancient groups, this approach provides limited resolution for reconstructing ancient demographic events (Bolnick 2011). Further, many Caribbean populations remain underrepresented in large catalogs of genomic variation. For instance, except for Barbadian Afro-Caribbeans, recently included in 1000 Genomes Phase 3, genetics research on most Lesser Antillean populations has been limited to uni-parental loci and low-density ancestry informative markers (Benn-Torres et al. 2012, Benn-Torres, Kittles, and Stone 2007, Benn-Torres, Stone, and Kittles 2013, Benn-Torres et al. 2015). The present research seeks to fill in these gaps through two approaches: an ancient DNA analysis of 60 individuals from three pre-contact archaeological sites in Puerto Rico

(A.D. 500–1300) and an analysis of genome-wide variants from 55 Afro-Caribbeans from five Lesser Antilles.

Specifically, in Chapter 2, I use the Illumina Infinium Multi-Ethnic Global Array to obtain autosomal genotypes from 55 self-identified Afro-Caribbeans from Grenada (n=6), St. Kitts (n=5), St. Lucia (n=15), Trinidad (n=19), and St. Vincent (n=10). I characterize patterns of genome-wide variation and ancestry in these groups and compare them to exising data from other recently admixed populations. I also perform two scans for genomic signatures of natural selection in these populations to identify candidate genes underlying local adaptation to Caribbean environments. In Chapter 3, I turn towards optimizing extraction methods for DNA recovery from ancient and degraded human and primate remains excavated in tropical environments from the Caribbean and East Africa. I compare the results of parallel extractions with dentine tissue across two methods and evaluate raw DNA yields as well as results obtained after mitochondrial enrichment capture. Lastly, in Chapter 4, I apply the methods developed in Chapter 3 as well as other protocols in a large-scale effort to recover and sequence complete mitochondrial genomes and partial autosomal genotypes from 60 human skeletal remains excavated from the pre-contact sites of Punta Candelero (n=19), Tibes (n=11) and Paso del Indio (n=30) in Puerto Rico. This analysis combines data generated by previous research with extant Native American groups and admixed Caribbean islanders to trace the origin and number of pre-contact migrations to Puerto Rico and examine the extent of genetic continuity between ancient and modern populations.

Overall, the findings from this research underscore the large impact of post-contact demographic shifts on Caribbean population history and illustrate how genomic

diversity has changed in this region over the last 7,000 years. In addition, this work increases the representation of admixed and diverse populations in available genomic datasets and has the potential to inform future functional and clinical genetics research with admixed Caribbean islanders.

# CHAPTER 2: UNDERSTANDING ADMIXTURE, GENETIC STRUCTURE AND ADAPTATION IN AFRO-CARIBBEANS FROM THE ENGLISH-SPEAKING LESSER ANTILLES

## 2.1 Abstract

Previous research with admixed Caribbean populations has shown that many islanders retain genomic variation from pre-contact indigenous groups, but also carry signatures of more recent admixture events fostered by European colonization and the African Slave Trade. However, most of this work has been conducted with Caribbean Latinos from the Greater Antilles, while populations from the Lesser Antilles remain underrepresented in large-scale genomics catalogs. This sampling gap precludes a complete understanding of the diversity of genomic variation across the Caribbean. Thus, to address this gap, this study analyzes autosomal SNP genotypes from 55 self-identified Afro-Caribbeans from five islands in the Lesser Antilles (LA): St. Kitts, St. Lucia, St. Vincent, Grenada, and Trinidad. I characterize patterns of genome-wide variation and ancestry in these groups and compare them to existing data from other recently admixed populations. I find that LA Afro-Caribbeans carry large proportions of African ancestry, but also have contributions from European, Native American, as well as South and East Asian populations to their autosomal genomes. This pattern is significantly different from that observed among admixed groups from the Greater Antilles. I further examine signatures of selection among LA Afro-Caribbeans to explore whether recent adaptation to the environmental pressures of the Caribbean may have shaped extant ancestry patterns. I find that LA Afro-Caribbeans carry signatures of selection at olfactory genes

that may be associated with the major histocompatibility complex on chromosome 6, among other loci. But whether these signatures stem from selection after continental admixture remains unclear. Findings from this project underscore the large impact of post-contact demographic processes on Caribbean population history and illustrate how genomic diversity has changed in this region since the initial occupation of the islands, 7,000 years ago. In addition, this work increases the representation of admixed and diverse populations in available genomic datasets and has the potential to inform future functional and clinical genetics research with admixed Caribbean islanders.

## 2.2 Introduction

Until recently, human genomics research was limited to a small sample of populations, most of which did not represent the full spectrum of human genetic diversity (Oh et al. 2015, Bustamante, De La Vega, and Burchard 2011, Popejoy and Fullerton 2016). To counter this bias, admixed groups such as Latinos and African Americans have recently been included in large surveys of genomewide diversity (1000 Genomes Project et al. 2015, Nelson et al. 2008, Moreno-Estrada et al. 2014, Bryc et al. 2015, Wang et al. 2008, Ruiz-Linares et al. 2014, Bryc, Auton, et al. 2010). Admixed Caribbean islanders in particular have received widespread attention as several recent studies have characterized their complex ancestry and demographic history (Gravel et al. 2013, Moreno-Estrada et al. 2013, Bryc, Velez, et al. 2010). However, the bulk of this work has focused on islanders from the Spanish speaking Greater Antilles, such as Puerto Ricans and Dominicans. Except for Afro-Caribbean individuals from Barbados, recently included in 1000 Genomes Phase 3, islanders from the Lesser Antilles remain

underrepresented in large scale catalogs of human genomic variation. Unfortunately, this sampling gap precludes a complete understanding of how differences in pre and post contact migration patterns, colonial experiences and local selective pressures have shaped genomic variation across the Caribbean. Beyond its anthropological or historical importance, such knowledge on genetic makeup is essential for informing efforts to identify phenotypic or medically relevant genetic variation in these populations, such as admixture mapping or genome wide association studies (Bustamante, De La Vega, and Burchard 2011, Kidd et al. 2012, Oh et al. 2015). Thus, to help build a more complete understanding of genetic diversity across the Caribbean, the present study characterizes genetic structure, ancestry and signatures of selection in Afro-Caribbeans from the English-Speaking Lesser Antilles.

The Caribbean islands, also known as the Antilles, form an archipelago that extends from North to South America across the Caribbean Sea. The islands are divided into three groups: the Bahamas, the Greater Antilles and the Lesser Antilles (LA) (Figure 1). The LA form an arc-like shape that begins with the Virgin Islands, located east of Puerto Rico, and ends with Trinidad and the Netherlands Antilles, along the South American coast (Rogozinski 2008). The first archaeological evidence for human occupation in the LA comes from the site of Banwari Trace in Trinidad, dated to 7,200 years before present (Harris 1973). After this date, archaeological evidence suggests that pre-contact migrations into the Antilles were numerous and continuous, and probably stemmed from multiple origin sites in South and Central America (Keegan 2013, Rouse 1992, Wilson 2007). At the time of European contact in 1492, multiple ethnic groups likely coexisted across the Caribbean basin (Wilson 2007, 1999). Spanish conquerors

identified the native inhabitants of the LA as the Caribs, and differentiated them from the Bahamian Lucayos and the Tainos and Ciboney/Guanahatabeys of the Greater Antilles. Whether this categorization corresponded to true ethnic, linguistic or cultural differences, however, remains a topic of extensive research (Petersen, Hofman, and Curet 2004, Curet 2014).

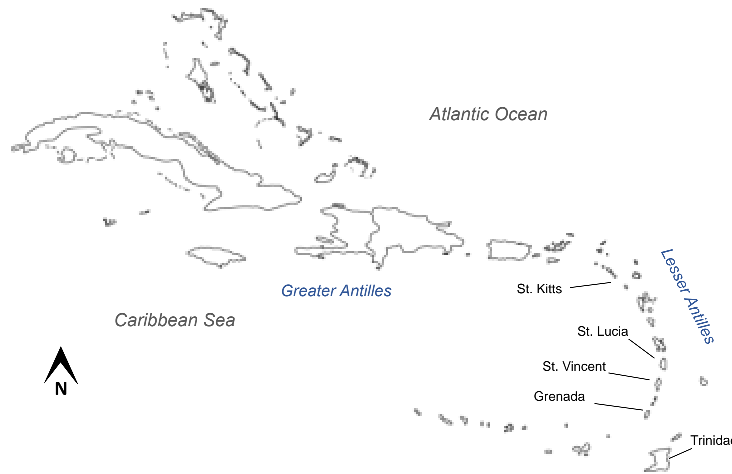


Figure 1. Map of the Caribbean Archipelago with islands included in present study.

European conquerors arrived in the Caribbean in the late 15$^{th}$ century but did not formally settle the LA until the 1600s. During the early colonial period the Spanish raided the LA islands to capture indigenous slaves for transport to the mining and agricultural colonies of the Greater Antilles and continental mainland (Rogozinski 2008, Anderson-Córdova 2005, Rivera-Pagán 2003). This practice, combined with overwork, abuse and introduction of new disease pressures led to dramatic reductions in the size of indigenous populations across the whole Caribbean basin. Demographic changes were also brought about by European colonists taking indigenous women as wives or concubines (Wilson 2007, Martin-Fragachan 1999). Indigenous groups in the LA,

sometimes in concert with runaways from the Greater Antilles, fiercely resisted European colonization (Anderson-Córdova 2005) surviving in maroon and isolated settlements until treaties with European powers granted them ownership of a subset of their former lands (Pérotin-Dumon 1999, Rogozinski 2008). Today, indigenous communities still live on the islands of St. Vincent (Garifuna), Dominica (Kalinago Reserve) and Trinidad (First Peoples Community Arima), while communities who self-identify as Native American exist in many other islands (Benn-Torres et al. 2015, Forte 2006).

Several colonial powers, including the French, Dutch and British formally settled the LA during the 1620s (Pérotin-Dumon 1999, Rogozinski 2008). Due to the decline of indigenous populations, indentured laborers from Northern Europe, as well as African slaves from several areas of West Africa, Central Africa and Mozambique were imported as labor for the growing plantation monocultures. Slave importation grew steadily during the next decades (Engerman and Higman 1997). By the time of emancipation in the 19th century, it is estimated the British had brought almost 2 million slaves, at a rate of about 50,000 slaves a year, to their Caribbean posessions (Pérotin-Dumon 1999, Rogozinski 2008). As a result, the demography of the islands shifted. European settlers and indigenous peoples became the minority while African slaves and their descendants became, on average, between 60 and 90% of the population. Moreover, sex ratios were skewed due to preferential importation of male slaves and indentured laborers (Engerman and Higman 1997, Rogozinski 2008). Liasons between mostly male Europeans and female slaves also led to the rise of a small creole population (Engerman and Higman 1997).

By the 19th century, the islands included in this research, St. Kitts, St. Vincent, St. Lucia, Grenada, and Trinidad were all under British control (Rogozinski 2008). When Britain abolished slavery in 1834, the subsequent labor shortage led to the revival of the indentured labour system. Between 1834 and 1917 over 500,000 indentured laborers, mostly from India, but also from China, Portugal and parts of Africa, came to work in the sugar plantations of the LA. Most laborers were transported to Trinidad and British Guiana although after completion of their contracts many migrated voluntarily to other islands such as St. Lucia. A large number of these migrants settled permanently, further transforming the demographic composition of the island populations (Engerman and Higman 1997, Laurence 2011, Lee-Loy 2004, Look Lai 2004). Today, the five islands included in this study are independent nations, with a population of approximately 1.65 million people (CIA 2013). Since the 1960s, citizens of the LA have migrated to the United States, the United Kingdom and other nations in steady numbers (Richardson 2004). As of the 2014 census, approximately 495,000 Lesser Antilleans, including people from Trindad, St. Vincent and Grenada, live in diasporic comunities in the United States (1.16% of the nation's estimated 42.4 million migrants) (Zong and Batalova 2016).

Although all Caribbean populations have similar histories of continuous migration and contact, different demographic patterns and colonial experiences have created heterogenous patterns of genetic structure and ancestry in this region (Schurr 2010). For example, Cubans and Puerto Ricans derive over 60% of their nuclear genomic ancestry from European populations, and have moderate proportions of Native American and African ancestry in nuclear loci (Gravel et al. 2013, Ruaño et al. 2009, Via et al. 2011, Marcheco-Teruel et al. 2014, Moreno-Estrada et al. 2013). But other islanders such as

Afro-Caribbeans from Trinidad & Tobago, Grenada, St. Vincent, and Dominica, derive most nuclear ancestry from Sub-Saharan Africa and carry very small proportions of European or Native American ancestry (Benn-Torres et al. 2008, Benn-Torres, Stone, and Kittles 2013, Miljkovic-Gacic et al. 2005). In addition, extensive sex-biased admixture has led to widely divergent mitochondrial and Y-chromosome ancestry proportions in these populations (Mendizabal et al. 2008, Benn-Torres, Kittles, and Stone 2007, Martínez-Cruzado et al. 2005, Simms et al. 2013).

Patterns of local adaptation to the selective pressures of the region also vary across Caribbean islanders. Previous work has scanned the genomes of admixed Caribbean Latinos searching for evidence that natural selection has operated since the onset of recent admixture, by locating genomic regions with excess or deficiencies in ancestry proportions relative to genomewide averages (Oleksyk, Smith, and Brien 2009, Tang et al. 2007). In Puerto Ricans, candidate selected regions with excess African and Native American ancestry have been identified on the human leukocyte antigen (HLA) region on chromosome 6 and in regions harboring olfactory gene clusters on chromosomes 8 and 12 (Tang et al. 2007). At least twelve additional ancestry deviations have also been detected among Puerto Ricans, Dominicans, Ecuadorians and Colombians, but these signtaures are not shared across all populations (Brisbin et al. 2012). Caribbean islanders were the first populations to experience the selective pressures introduced by European contact, such as exposure to disease and to novel or changing tropical environments (Cook 1998, Salvaggio 1992). Thus determining whether local adaptation has occurred in these groups and which genomic loci are involved can contribute to understanding how natural selection shapes ancestry patterns in recently

admixed populations and can lead to identification of phenotypically important rare genetic variants (Tishkoff 2015).

At present, human genetics research in the LA has largely focused on uni-parental loci (mitochondrial DNA or Y-chromosome) and low-density autosomal ancestry informative markers (Benn-Torres, Kittles, and Stone 2007, Benn-Torres et al. 2015, Benn-Torres, Stone, and Kittles 2013, Toro-Labrador, Wever, and Martínez-Cruzado 2003, Benn-Torres et al. 2012, Miljkovic-Gacic et al. 2005). Given the recent history of sex-biased admixture in these populations and the limited capacity of uniparental markers to provide information about autosomal ancestry (Emery et al. 2015), these studies provide an important but incomplete picture of genetic diversity in the LA. Recent efforts to include admixed Caribbean groups in large surveys of genomewide diversity (Bryc, Velez, et al. 2010, Gravel et al. 2013, Moreno-Estrada et al. 2013, Montinaro et al. 2015) and in genomic scans for selection (Pybus et al. 2013, Deng et al. 2016) have begun to address this gap, but so far have included few individuals from the English-Speaking Antilles. In this investigation, I expand on this body of previous research by analyzing nuclear genotypes from 55 Afro-Caribbean individuals from the islands of Grenada (n=6), St. Kitts (n=5), St. Lucia (n=15), St. Vincent (n=10), and Trinidad (n=19). With these data I seek to address the following questions:

(1) Do genomewide ancestry and structure patterns differ across Afro-Caribbeans in the LA and between the LA and the Greater Antilles?
(2) What is the influence of selection on the distribution of ancestry in admixed Caribbean populations?

To answer these questions, I analyzed a combined dataset of newly generated genomewide single nucleotide polymorphism (SNP) data from Lesser Antillean Afro-Caribbean groups and published genotypes from seven admixed Caribbean Latino populations: Cuba, Haiti, Dominican Republic, Puerto Rico, Barbados, Honduras, and Colombia (Moreno-Estrada et al. 2013, 1000 Genomes Project et al. 2015). I also collected reference population genotypes from published sources (Reich et al. 2012) and publicly available panels such as the 1000 Genomes Phase 3 (1000 Genomes Project et al. 2015) and the Human Genome Diversity Panel (HGDP) (Cann et al. 2002, Rosenberg et al. 2002). To answer my first research question, I applied multivariate and unsupervised clustering methods to characterize population structure and global ancestry among LA Afro-Caribbeans. I also used local ancestry and deconvolution methods to determine the continental and sub-continental origin of admixed haplotypes in these individuals. I then compared ancestry proportions between autosomal and X-chromosome loci to assess whether sex-biased ancestry had a large effect in the studied populations. To address my second research question, I employed long-range haplotype selection scans and ancestry deviation scans to identify signatures of selection in the LA Afro-Caribbean population. I then used gene ontology methods to predict the function of selected candidate regions. Throughout the study, I compare my results to those obtained by previous researchers examining ancestry, population history and selection in Caribbean Latino populations and other recently admixed human groups.

Complex patterns of admixture since European contact were identified among LA Afro-Caribbeans, including a strong signature of sex-biased mating and inputs from at least five continental populations to their autosomal ancestry. I also observed differences in

admixture proportions across Caribbean basin populations which likely stem from differing experiences with European colonization, the African Slave Trade and modern economic development. Lastly some evidence was found for the action of natural selection in admixed Afro-Caribbeans, but reliably identifying whether the signals arose before or after contact period admixture is difficult. To my knowledge, this study is the first analysis of genome wide ancestry and selection patterns in Afro-Caribbean population in the Lesser Antilles.

## 2.3 Materials and Methods

### 2.3.1 *Sample Collection and Genotyping*

Samples included in this study were previously collected by co-author Dr. Jada Benn Torres among unrelated, self-identified Afro-Caribbean individuals, 18 years or older, who were born in, or had at least one parent or three grandparents from the English-speaking Lesser Antilles (LA). Original sample collection was conducted between 2004 and 2005 with informed consent and Institutional Review Board (IRB) approval granted by University of New Mexico and local ministries of health from the respective islands. In 2014, a second IRB approval (STUDY00001580) and a Material Transfer Agreement were granted for the present study by the respective universities (Appendix A). Only participants that provided consent for future study were included in these analyses. As described in Benn-Torres, Kittles, and Stone (2007), DNA samples were originally collected with buccal swabs and extracted using standard phenol-chloroform methods. Between June 2015 and May 2016, 88 de-identified DNA extracts from the following island populations were transported to ASU from Dr. Benn-Torres'

laboratory: Grenada (n=9), St. Kitts (n=8), St. Lucia (n=24), Trinidad (n=27), and St. Vincent (n=20) (Table SM1).

Extracts were quantified with the Qubit 2.0 Broad Range assay (Simbolo et al. 2013). A subset of samples with low DNA concentrations were whole genome amplified (WGA) with the Qiagen REPLI-G Mini Kit following manufacturer's instructions. WGA products were purified using Qiagen Qiaquick columns and re-quantified with the Qubit. Eleven samples with low post-WGA DNA concentrations or with less than 10 uL remaining volume were excluded from further processing at this stage. The 77 remaining samples were subjected to DNA genotyping on the Illumina Infinium ® Multi-Ethnic Global Array (MEGA$^{EX}$). This array includes over 2 million single nucleotide polymorphisms (SNPs) genome-wide distributed and was specifically designed to reduce ascertainment bias in ethnically diverse populations (Bien et al. 2016, Illumina 2015). For each of the 77 samples, between 300-500 ng of DNA were submitted for genotyping. Seventeen of these samples consisted of DNA extracts without amplification and the remaining 60 consisted of whole genome amplified DNA. In addition, 13 WGA samples were genotyped in duplicate, increasing the total number of samples placed on the array to 90 (Table SM2). Genotyping was performed at the University of Miami Miller School of Medicine, John P. Hussman Institute for Human Genomics. Raw genotype data will be available in the NCBI dbGAP database through data access agreements after publication.

2.3.2 *Data Curation and Quality Control*

All 90 samples were successfully genotyped for 2,036,060 unique SNP markers on the MEGA$^{EX}$ array. Genomic markers and coordinates were mapped to human

genome build GRCh37 (hg19). Data quality assessments and input filtering were performed following guidelines as listed in (Anderson et al. 2010) and using Plink 1.9 (Chang et al. 2015). Data were transformed from Illumina Final Report format to Plink format using custom bash scripts. After filtering for multi-allelic markers, 1,933,206 bi-allelic sites remained. 83,397 SNPs mapped on the reverse strand were identified using snpflip (https://github.com/endrebak/snpflip) and flipped to the forward strand with Plink 1.9. The complete dataset was restricted to autosomal SNPs. Per-individual quality filtering resulted in 33 samples being excluded due to excess heterozygosity values (which may be indicative of genotyping errors), excess missing genotype rates (>5% missing data) and/or high identity-by-descent values indicative of hidden relatedness or sample duplication (Figure SM1). When an individual was genotyped more than once, the duplicate with less missing data was kept for further analysis. No individuals were eliminated due to discordant sex information.

Per-marker quality filtering resulted in exclusion of 86,460 SNPs with >10% missing data, 11,179 SNPs under Hardy-Weinberg equilibrium (HWE) threshold P< $1x10^5$, and 779,643 SNPs with minor allele frequency (MAF) <1%. Lastly, 45,160 duplicate markers or markers without rs identifiers were also removed. The final analysis ready dataset included N=55 unrelated, unique individuals from the LA and 1,010,700 autosomal SNPs, with an average heterozygosity of 0.236 and an average genotyping call rate of 0.992. Final sample breakdown per population was as follows: Grenada (n=6), St. Kitts (n=5), St. Lucia (n=15), Trinidad (n=19), and St. Vincent (n=10).

2.3.3 *Data Merging and Integration*

To conduct comparative analyses, the quality filtered LA dataset was merged with several continental reference panels including: world populations from 1000 Genomes Phase 3 (1KG) (1000 Genomes Project et al. 2015) and HGDP (Cann et al. 2002, Rosenberg et al. 2002), Native American populations from Reich et al. (2012), and admixed Caribbean populations from Moreno-Estrada et al. (2013) (Table SM3). The data from Reich et al. (2012) had been previously masked such that segments of potential non-Native American ancestry were set as missing and related individuals were eliminated. This approach prevents recent admixture from confounding ancestry estimation performed with this panel but reduces the total amount of SNPs in the dataset. The samples from Moreno-Estrada et al. (2013) consisted of family trios, therefore all offspring individuals were excluded from analysis. The HGDP panel was trimmed to a standardized subset of 940 unrelated individuals (Rosenberg 2006) using scripts from the Terastructure package (https://github.com/StoreyLab/terastructure/blob/master/scripts/data/HGDP_text_to_tped.py).

For reference datasets mapped to previous human genome builds, genomic coordinates were lifted to GRCh37 using the UCSC Genome Browser liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) and the liftOverPlink wrapper (https://github.com/sritchie73/liftOverPlink). For all reference datasets, SNPs were flipped to the forward strand using methods as described above. For each merge, the intersection of common SNPs between the LA dataset and each of the pertinent reference panels was identified using custom R scripts (R Core Team 2016). The datasets were

merged at these intersections using Plink 1.9 and per marker quality filters (MAF>10%, missing data <5%, HWE P< $1x10^{5}$) were reapplied.

Multiple combinations of merged datasets were produced due to the differing requirements of each analysis and the different SNP densities obtained (Table SM4). Merged dataset 1 (119,277 SNPs after quality filtering) included the five Lesser Antillean groups and all global reference populations (n=3,529). Merged dataset 2 (36,928 SNPs after quality filtering) consisted of all populations in dataset 1 plus the Caribbean populations from Moreno-Estrada et al. (2013) (n=3,688). The latter has much less SNP density due to the poor intersection between Illumina and Affymetrix arrays. Subsets of these two datasets were used for ADMIXTURE and principal components analysis (PCA). Merged dataset 3 (852,887 SNPs after quality filtering) included Lesser Antillean populations, and four populations from the 1000 Genomes Phase 3 Panel: Yoruba (YRI), Utah residents of European ancestry (CEU), Han Chinese (CHB), Tamil (ITU) (n=467). These were chosen based on the second run of ADMIXTURE results as putative representatives of ancestral populations (see next section). This dataset also had the highest SNP density due to good intersection between the $MEGA^{EX}$ genotyped SNPs and 1000 Genomes sequence data. Merged dataset 4X included the same populations as merged dataset 3 but was restricted to just female individuals and X chromosome loci (30,659 X chromosome SNPs after quality filtering, n=229). These two datasets were used to investigate signatures of sex-biased admixture (see section on X chromosome analyses). Merged dataset 3 was also used for natural selection scans. Native American populations were not included in these two datasets because no masked data were available for chromosome X from the Reich et al. (2012) reference panel. Lastly, merged

dataset 5 (119,277 SNPs after quality filtering) was a subset from dataset 1. It included the five Lesser Antillean groups and representative populations for each of the five continental regions identified in the ADMIXTURE analyses. These included the YRI, CEU, CHB and ITU populations referenced above, and 105 indigenous individuals from Central and South America from the Reich et al. (2012) reference panel (n=572). Multiple individuals from several populations were selected so that the sample size of the Native American reference group would be comparable to those of the other four populations. This dataset was used for local ancestry inference of autosomal genotypes and for natural selection scans.

2.3.4 *Global ancestry estimation and Principal components analyses*

To conduct multivariate and unsupervised clustering analyses, merged datasets 1 and 2 were thinned for linkage disequilibrium (LD) using the indep-pairwise option in Plink 1.9 (--indep-pairwise 50 10 0.1). This command marks for removal SNPs with a pairwise $r^2 > 0.1$ within a 50 SNP sliding window, shifted forward by 10 SNPs at a time. After the pruning step, 44,219 SNPs remained in merged dataset 1, and 23,636 SNPs remained in merged dataset 2. PCA was performed on both pruned datasets using smartpca from the EIGENSOFT 6.0.1 package (Patterson, Price, and Reich 2006).

Unsupervised global ancestry clustering analyses were conducted on the merged autosomal genotype datasets using ADMIXTURE v1.22 (Alexander and Lange 2011, Alexander, Novembre, and Lange 2009). This program estimates global ancestry proportions by testing data fit to an admixture model where K user defined populations have contributed to an admixed individual's genome. The analysis is 'unsupervised'

because population labels are not assigned a priori. Instead groupings are determined from the data itself (Liu et al. 2013). The first ADMIXTURE run, performed with a subset of merged dataset 1, included the LA populations, and a combined subset of HGDP and 1KG populations from five continental regions: Africa (Esan, Yoruba, Mende, Gambian), Europe (British, Utah, Iberian Spanish, French, Tuscan), Central-South Asia (Gujarati, Tamil, Sindhi, Bengali), East Asia (Dai, Han), the Americas (Maya, Mixe, Kaqchikel, Colombian Piapoco, Surui, Karitiana, Central Amerind, Ge Pano Carib, Chibchan Paezan, Equatorial Tucanoan, Andean) (Table SM3). The second run was performed with a subset of merged dataset 2 and it included the same populations listed above, plus admixed Caribbean individuals from Moreno-Estrada et al. (2013) and 1KG (Colombians, Hondurans, Cubans, Dominicans, Haitians, Puerto Ricans, Afro-Caribbean Barbarians). Both ADMIXTURE analyses were performed exploring clusters from K=2 to K=15. The lowest cross-validation error was observed at K=10 for dataset 1 and K=11 for dataset 2 (Figure SM2). Clustering models above K=5 resulted in within continent substructure.

Autosomal global ancestry proportions were compared between populations with the pairwise Wilcoxon rank sums test as in Homburger et al. (2015). This test is a non-parametric alternative to a t-test that assesses whether significant differences exist between two distributions. It does not require data points to be normally distributed, and it is robust to skews and the presence of outliers (Moore, McCabe, and Craig 2009). In this case the null hypothesis tested was that the distribution of a given global ancestry proportion was identical across all populations. Assuming a model of K=5 ancestral components, differences in ancestry proportions from five continental regions were

tested: Africa, Europe, South Asia, East Asia and Americas. The test was applied to detect significant differences in ancestry proportions between all LA populations from merged dataset 1, and between all Caribbean populations included in merged dataset 2. It was conducted using the pairwise.wilcox.test function in R with a significance cutoff of $P<0.05$ and Bonferroni correction for multiple testing. The distribution of ancestry proportions across populations was visualized with boxplots.

Based on the first two ADMIXTURE runs four reference populations were chosen for inclusion in merged datasets 3 and 4X. These datasets included the admixed Lesser Antilleans and four global reference populations chosen to represent putative African, European, South Asian and East Asian ancestral components (YRI, CEU, ITU, and CHB). In addition, 105 Native American individuals from several Central and South American populations were selected for inclusion in merged dataset 5. Stacked bar plots visualizing ADMIXTURE results were produced with pong (Behr et al. 2016). Figure labels and colors were modified with Adobe Illustrator.

### 2.3.5 *Phasing and Local Ancestry Estimation*

Merged datasets 3, 4X and 5 were phased with SHAPEIT2 (Delaneau, Marchini, and Zagury 2012). Before phasing, centimorgan (cM) positions for all markers were added to the merged dataset with Plink 1.9 (using option --cm-map) and data were separated by chromosome. The 1000 Genomes Phase 3 genetic map was used a reference haplotype panel for phasing and cM mapping (https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html). Local ancestry estimation was conducted using RFMix v1.5.4 (Maples et al. 2013). RFMix uses a discriminative, random forest approach to infer local

ancestry of chromosomal segments from phased genotypes. The program also performs phase correction and improves accuracy in ancestry calls through an iterative expectation-maximization (EM) algorithm. The program was run separately per each chromosome using the PopPhased option. Window size was set at 0.2 cM and the number of EM iterations was set to 2. Local ancestry estimation was performed for merged datasets 3, 4X and 5. Phased genotypes from CEU, YRI, ITU, CHB, and the 105 Native Americans were designated as reference populations and phased genotypes from Grenada, St. Kitts, St. Lucia, St. Vincent and Trinidad were considered as admixed. Program output included ancestry calls (Viterbi files) and posterior probabilities of each reference ancestry (Forward-Backward files) per SNP. Ancestry call cutoffs were determined with a 0.9 posterior probability threshold as recommended in Kidd et al. (2012). Local ancestry per individual was visualized in color-coded karyogram plots produced with Python 2.7.11. Lastly, global ancestry proportions per each admixed individual were calculated from RFMix local ancestry proportions and visualized in stacked bar plots. RFMix input files, output files and karyogram plots were processed with publicly available Python scripts written by A.R. Martin (https://github.com/armartin/ancestry_pipeline) and global ancestry bar plots were produced with custom scripts written in R.

2.3.6 *X chromosome analyses*

To identify signatures of sex-biased admixture in the LA, ancestry proportions on the X chromosome were compared to those on the autosomes. X chromosome genotype data was obtained and merged using the same methods as discussed above for the CEU,

YRI, ITU, CHB and LA populations. Masked Native American X chromosome genotypes were not available in the Reich et al. (2012) reference panel, therefore this ancestry component was excluded from this analysis. In addition, only LA females were kept to ensure comparisons were made exclusively between diploid chromosomes. Quality filters were re-applied after merging intersecting SNPs and excluding all males (Table SM4). PCA and ADMIXTURE analyses were conducted on merged dataset 4X after LD pruning (6,175 SNPs remained in pruned dataset) using the same methods as detailed above. Clustering models from K=2 to K=6 were explored, and the lowest cross-validation error was observed at K=3. Phasing and local ancestry estimation were performed on the non-LD pruned dataset using the same methods as above.

Chromosome X global ancestry proportions for each admixed female were calculated from RFMix local ancestry estimates. These were compared to autosomal global ancestry proportions obtained through the same methods, for the same individuals, with the Wilcoxon signed-ranks test as in Moreno-Estrada et al. (2013). The test was applied as described above (see Global ancestry estimation and Principal components analyses section), except the data were treated as paired. The null hypothesis tested was that the distribution of ancestry proportions on the autosomes was identical to that on the X chromosome for each ancestry component (Africa, Europe, South Asia, East Asia). The test was applied to all LA females (n=23) to detect overall trends, and to females from St. Lucia (n=11) and St. Vincent (n=7) to detect differences between populations. Grenada, St. Kitts and Trinidad were excluded from the population specific tests because less than four females were successfully genotyped from each. The test was conducted using the wilcox.test function in R with a significance cutoff of $P<0.05$. Differences in

the distribution of ancestry proportions between the X chromosome and the autosomes were plotted as box and barplots.

2.3.7 *Scans of natural selection: Haplotype based measures*

To identify candidate regions under positive selection in the admixed Lesser Antillean populations the integrated haplotype score (iHS) and the extended cross-population haplotype homozygosity measure (XP-EHH) were calculated. These test statistics detect long range haplotypes which arise during positive selection when advantageous alleles are in LD with neighboring variants (Voight et al. 2006, Sabeti et al. 2007). iHS is a within-population test that can detect incomplete selective sweeps resulting in intermediate to moderate frequency alleles that have not yet reached fixation. Calculation of iHS scores estimates the loss of diversity in haplotypes surrounding a putative selected SNP. iHS values can be positive or negative, depending on whether selection is occurring on the derived or ancestral allele, respectively. However, in practice most analyses are conducted on the absolute value of the iHS score for simplicity (Voight et al. 2006). Although this measure is calculated per SNP, grouping iHS scores into 100-200 kb windows of consecutive SNPs can increase detection power because selective events tend to produce clusters of extreme scores (|iHS > 2|) (Pickrell et al. 2009, Voight et al. 2006). Windows with high frequencies of top 1% outlier scores (values outside of the 99% confidence interval) are considered candidate regions under positive selection (Voight et al. 2006).

In contrast, XP-EHH is a between-population test that detects selective sweeps where alleles have risen to extremely high frequencies or complete fixation in one

population, but remain polymorphic in the other (Sabeti et al. 2007). XP-EHH values can be positive or negative, indicating selection has occurred in tested populations A or B, respectively. As with iHS, extreme XP-EHH scores are examined as absolute values and can be grouped into windows to achieve higher power for detecting candidate regions with long stretches of extreme scores (Vitti, Grossman, and Sabeti 2013). In humans, iHS and XP-EHH tests can detect recent selective sweeps within the last 30,000 years (Oleksyk, Smith, and Brien 2009).

Phased genotypes from merged dataset 3 were used as input for the iHS and XP-EHH selection scans. This dataset was chosen because it had the highest SNP density. Ancestral and derived allele states were retrieved from the inferred human ancestor sequence (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/human_ancestor_GRCh37_e59.tar.bz2) using scripts from the Selection Tools pipeline (Cadzow et al. 2014). Simulations have shown that to maintain statistical power, iHS and XP-EHH need a minimum sample size of 40 and 20 chromosomes (20 and 10 diploid individuals), respectively (Pickrell et al. 2009). Therefore, all genotyped LA populations were grouped together for these analyses for a total sample size of 55 diploid individuals (110 chromosomes). Although this approach precludes detection of selected candidate regions that may differ between populations, it ensures robust detection of true positive selection signals common to all groups and limits detection of false positives.

Using Selscan (Szpiech and Hernandez 2014), unstandardized iHS scores were calculated for the phased data from the LA Afro-Caribbeans and for each of the four reference populations in merged dataset 3 (YRI, CEU, ITU, CHB). The program was run

with default parameters, including a MAF cutoff threshold of < 0.05 and critical value threshold of | iHS > 2 |, as recommended by (Voight et al. 2006). Test statistic values were standardized using default parameters with the *norm* tool provided within Selscan. Standardization normalizes iHS scores with mean = 0 and standard deviation = 1. Standardized iHS scores were then grouped into 100 kb windows with a minimum threshold of 10 SNPs per window. The fraction of scores above the critical value was calculated for each window per chromosome, for all tested populations. Manhattan plots of standardized iHS scores were produced using the qqman package in R (Turner 2014).

For the XP-EHH statistic, four pairwise comparisons were calculated comparing the LA group to each of the four reference populations, respectively. Score calculation, standardization and binning across 100 kb windows were performed in Selscan and norm with default parameters, in the same way as described above. For iHS and XP-EHH, variability in the number of variants per 100 kb window was accounted for by binning windows with similar number of SNPs in increments of 10 SNPs as in Pickrell et al. (2009) (Figures SM3 and SM4). For each population, the 100 kb windows with the highest proportions of extreme iHS or XP-EHH scores (in the top 1% and 5%) were considered candidate regions under selection (Kilman 2016).

2.3.8 *Scans of natural selection: Ancestry deviations*

A second selection scan was conducted to identify candidate regions under selection that have arisen since the onset of admixture in Lesser Antillean populations. This approach is similar to admixture mapping in that it searches the genome of an admixed population for deviations from genome wide ancestry proportions due to excess

or deficiency of a given ancestry component. These deviations, also known as ancestry skews, may stem from selection acting on the frequency of alleles that are positively selected in a population after an admixture event (Tang et al. 2007, Oleksyk, Smith, and Brien 2009). RFMix local ancestry estimates for merged datasets 3 and 5 were used as input for the ancestry deviation scan. Based on the results of previous global ancestry analyses (see Global ancestry estimation and Principal components analyses section), a model of five-population admixture was assumed using YRI, CEU, ITU, CHB and 105 Native Americans as references. As with the long-range haplotype analyses, admixed individuals from all five LA populations were grouped together for the first round of ancestry deviation tests. However, the test was also run for each island population independently to explore whether grouping them together affected statistical power and detection of false positives.

Deviations in locus-specific ancestry per each ancestral component were calculated as in (Chimusa et al. 2015) and (Tang et al. 2007) by subtracting local ancestry proportions per marker from genome-wide global ancestry proportions as follows:

$$\delta_k^m = \left(\frac{1}{N}\sum \varphi_k^{i,m}\right) - \alpha_k = \overline{\varphi}_k^m - \alpha_k$$

N is the number of individuals in the admixed population sample. For each ancestral component $k \in \{1, \ldots, K\}$, $\alpha_k$ is the genome-wide global ancestry proportions derived from local ancestry estimates and $\varphi_k^{i,m}$ is the locus-specific ancestry directly estimated by RFMix of each individual i at SNP $m \in \{1, \ldots, M\}$. $\overline{\varphi}_k^m$ is the averaged locus-specific ancestry for all sampled individuals at SNP m. The deficiency or excess of ancestry $\delta_k^m$ can be approximated as a normal distribution with mean = 0 and standard deviation = 1

(Daya et al. 2014, Deng et al. 2016, Chimusa et al. 2015). As recommended by (Bhatia et al. 2014), a strict threshold for significance was implemented to avoid detection of false positives. Thus, only genomic segments with normalized deviation scores above four standard deviations from the mean were considered as candidate selected regions. All calculations and figures depicting the distribution of ancestry proportions and deviations across the genome were done using custom scripts in R.

2.3.9 *Gene ontology and overrepresentation analyses*

To identify genes within candidate selected regions identified with the genome-wide selection scans, genomic annotations were examined with the Ensembl genome browser for human build GRCh37 (Yates et al. 2015). Specifically, unique Ensembl identifiers were obtained for all genes within candidate regions using the BiomaRt R interface for retrieving Ensembl annotations (Durinck et al. 2005, Durinck et al. 2009). The list of Ensembl identifiers was used for gene function analysis with the PANTHER protein annotation through evolutionary relationship database v11.1. This database retrieves functional classification of genes by family, protein class, biological process or molecular function, among other categories (Mi et al. 2016). PANTHER was also used to conduct a statistical overrepresentation test of the biological processes of top candidate genes. This is a binomial test which compares the input gene list to a reference list (in this case 20,972 known annotated genes in the human genome reference) to determine whether a gene class is over or under represented. The null hypothesis is that genes in the input list are sampled from the same population as genes in the reference and therefore the probability of observing a given biological process gene in the input list should be

equal to that of the reference. Test significance cutoff was $P<0.05$ and p-values were adjusted for multiple testing with the Bonferroni correction (Mi et al. 2013).

2.3.10 *Computational resources and R packages*

This research was completed using resources from the ASU High Performance Computing Saguaro environment. All custom scripts written for this project will be available on GitHub after publication. All plots and figures were generated in R 3.2.4 using the ggplot2 (Wickham 2009) and reshape 2 (Wickham 2007) packages or with R base graphics.

## 2.4 Results

2.4.1 *Population Structure and Global Ancestry in the Lesser Antilles*

Multivariate PCA and unsupervised global ancestry analyses were used to characterize population structure and ancestral components of LA Afro-Caribbean populations. Figure 2 shows the distribution in PC space (PC 1 vs 2) of individuals included in the first round of analysis: LA and global populations from seven world regions (merged dataset 1). Most LA individuals cluster with African populations, but there is variation in clustering patterns across island groups and even between individuals from a same population. For instance, many Trinidadians cluster between African, European, Central-South Asian and Native American reference populations, suggesting they carry ancestry from non-African sources. At least one individual from St. Lucia and two from St. Vincent and Grenada also fall between continental populations. However, within island variation is evident. Some Trinidadians cluster tightly with individuals from

St. Kitts, Grenada and St. Vincent, all of whom also group closely with African populations. These clustering patterns are largely preserved at higher PC components (Figure SM5).

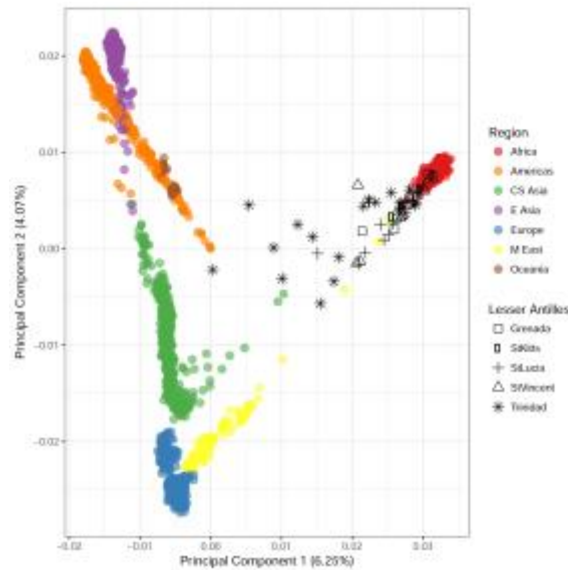


Figure 2. PCA performed with 44,219 autosomal SNPs intersected between Lesser Antilles genotyped samples and reference populations from Africa, Europe Central-South Asia, East Asia, Oceania and the Americas collected from the literature. PC 1 and 2 contain 10.32% of total variation.

Results of a second round of PCA including additional Caribbean basin populations (merged dataset 2) are shown in Figure 3. As expected, many Afro-Caribbean LA individuals cluster closely with Afro-Caribbean Barbadians, Haitians and continental African populations. This suggests that, on average, sampled LA groups carry more African ancestry than Puerto Ricans, Hondurans, Colombians and some Dominicans and Cubans, who cluster closer to European and Native American reference populations. However, at least one Trinidadian individual clusters with a group of Colombian and Honduran individuals who fall close to Native American reference populations. Clinal distributions of individuals across PC space are observed throughout

the Caribbean sample, suggesting that variation in ancestry components within and between island populations are common throughout the Caribbean basin. Additional plots with higher PCs for both rounds of PCA are shown in Figure SM6

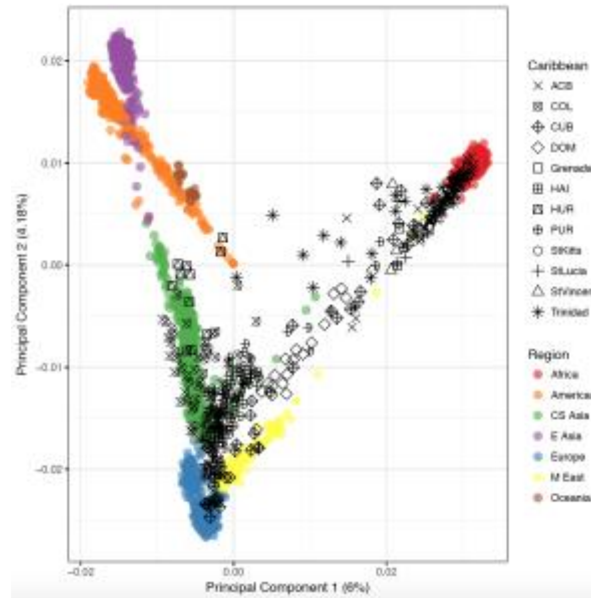


Figure 3. PCA performed with 23,636 autosomal SNPs intersected between Lesser Antilles genotyped samples, reference populations from Africa, Europe Central-South Asia, East Asia, Oceania and the Americas, and admixed Caribbean populations. PC 1 and 2 contain 8.18% of total variation. COL = Colombia, HUR = Honduras, CUB = Cuba, HAI = Haiti, DOM= Dominican Republic, PUR = Puerto Rico, ACB = Afro-Caribbean Barbados.

ADMIXTURE was used to estimate global ancestry proportions in the LA sample and to explore the fit of admixture models for several values of K ancestral populations. The analysis was run twice. The first run focused on characterizing global ancestry patterns in the LA with respect to world populations from Africa, Europe, Central-South Asia, East Asia and the Americas (Figure 4). The second run added admixed Caribbean basin populations to compare the distribution of global ancestry proportions across the region (Figure 5) (see *Global ancestry estimation and Principal components analyses section* in Methods section and Table SM3 for details*)*. For both runs clustering models

from K=2 to K=15 were explored and the lowest cross-validation (cv) error was observed at K=10 and K=11, respectively (Figures SM7 & SM8). However, clustering models above K=5 resulted in within continent substructure and did not further differentiate major continental clusters within the admixed genotypes. Therefore, a model where a maximum of K=5 continental populations contributed to genomic admixture was assumed to best explain overall ancestry patterns across the LA.

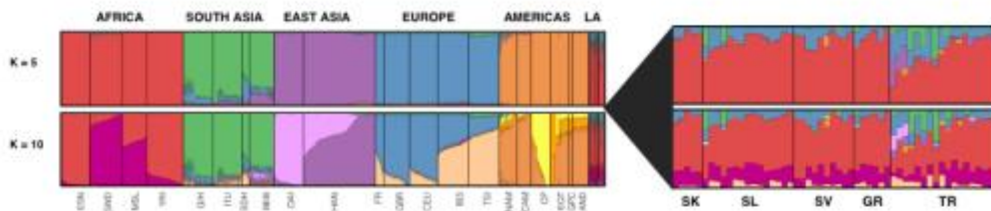


Figure 4. ADMIXTURE clustering performed with 44,219 autosomal SNPs common to Lesser Antilles genotyped samples and reference populations from the Caribbean, Africa, Europe South Asia, East Asia and the Americas (merged dataset 1). See Table SM3 for abbreviations.

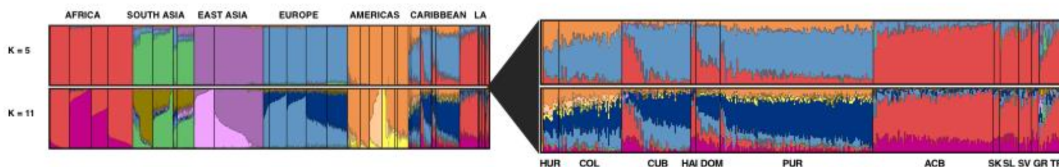


Figure 5. ADMIXTURE Clustering performed with 23,636 autosomal SNPs common to Lesser Antilles genotyped samples and reference populations from the Caribbean, Africa, Europe South Asia, East Asia and the Americas (merged dataset 2). See Table SM3 for abbreviations.

All LA populations show a mixture of African, European and Native American ancestry (Table 1), and some populations also carry contributions from South and East Asian sources. Although the African component is the largest in all groups, ancestry

proportions vary widely between populations (65% to 89%). Native American ancestry is the smallest global ancestry component observed across all groups (<2.5%), sometimes found in lower proportions than East or South Asian ancestry. Of all the LA populations, Trinidad stands out as having the most complex global ancestry patterns and the most variation in ancestry proportions for each of the five continental components (Figure 6). Trinidadian Afro-Caribbeans carry the highest proportions of East and South Asian ancestry, and the lowest proportion of African ancestry of all the LA populations. Afro-Caribbean populations from Trinidad and St. Vincent, two islands with recognized Native American communities, carry the highest proportions of Native American ancestry across the LA sample. However, when tested with a pairwise Wilcoxon rank test with Bonferroni correction these between population differences were not statistically significant (Table SM5). It must be noted however that the low number of individuals sampled from Grenada and St. Kitts may introduce some sampling bias and may mask the presence of more complex admixture patterns in these populations.

Table 1. Autosomal global ancestry estimates for Lesser Antilles Afro-Caribbean populations using RFMix and ADMIXTURE.

| Population | N | Estimate | AFR | EUR | AMR | SAS | EAS |
|---|---|---|---|---|---|---|---|
| Grenada | 6 | RFMix | 87.52% | 10.10% | 0.44% | 1.80% | 0.14% |
| | | ADMIXTURE | 86.23% | 11.36% | 0.68% | 1.22% | 0.52% |
| St. Kitts | 5 | RFMix | 90.51% | 9.18% | 0.09% | 0.17% | 0.05% |
| | | ADMIXTURE | 89.27% | 9.74% | 0.25% | 0.32% | 0.42% |
| St. Lucia | 15 | RFMix | 82.58% | 13.45% | 0.42% | 3.28% | 0.27% |
| | | ADMIXTURE | 81.00% | 14.36% | 0.68% | 3.47% | 0.50% |
| St. Vincent | 10 | RFMix | 84.68% | 12.72% | 1.68% | 0.75% | 0.18% |
| | | ADMIXTURE | 83.22% | 13.54% | 1.89% | 1.10% | 0.24% |
| Trinidad | 19 | RFMix | 67.67% | 16.78% | 1.77% | 8.38% | 5.40% |
| | | ADMIXTURE | 65.98% | 17.59% | 2.03% | 8.21% | 6.19% |
| **All Lesser Antilles** | **55** | **RFMix** | **82.59%** | **12.45%** | **0.88%** | **2.87%** | **1.21%** |
| | | **ADMIXTURE** | **77.54%** | **14.58%** | **1.33%** | **4.14%** | **2.41%** |

Abbreviations: AFR = African ancestry, EUR = European ancestry, AMR = Native American ancestry, SAS = South Asian ancestry, EAS = East Asian ancestry.

With the addition of seven admixed Caribbean populations in the second run of ADMIXTURE analysis, stark differences become apparent in the distribution of global ancestry proportions across the region (Figure 5). Even with the lower SNP density of merged dataset 2, overall admixture patterns are replicated from the previous run and K=5 remains the best model for maximum continental contributions. As expected based on the results of PCA, ancestry proportions were most similar between Afro-Caribbean Barbadians (ACB), Haitians and LA populations who carried much higher African ancestry, and much lower Native American ancestry than Hondurans, Colombians, Dominicans and Puerto Ricans. These overall similarities and differences in ancestry proportions were found to be statistically significant at $\alpha$=0.05 for multiple comparisons between Caribbean populations (Figure 7 and Table SM6). Cubans were the only Greater Antilles population with a small East Asian component, although in much lower proportions than Trinidadians.

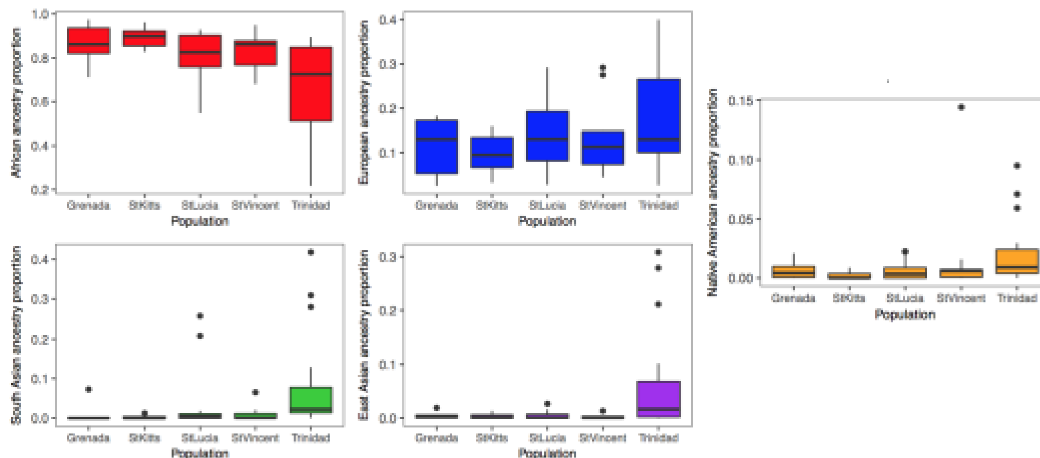


Figure 6. Distribution of continental ancestry proportions among Lesser Antilles Afro-Caribbeans generated with ADMIXTURE analysis at K=5 with 44,219 autosomal SNPs (merged dataset 1).

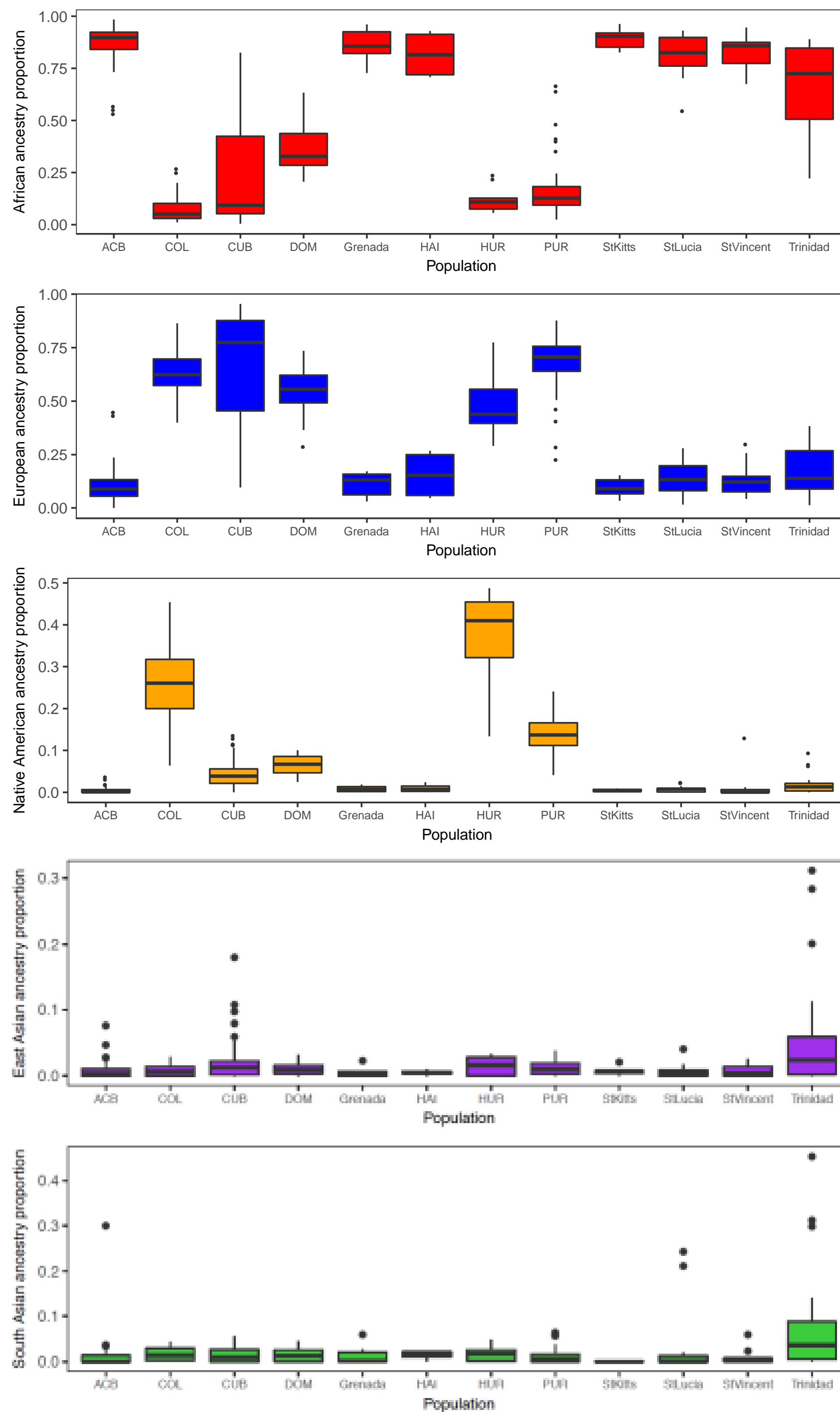


Figure 7. Distribution of continental ancestry proportions across 12 admixed Caribbean populations generated from ADMIXTURE analysis at K=5 with 23,636 autosomal SNPs (merged dataset 2).

With the addition of seven admixed Caribbean populations in the second run of ADMIXTURE analysis, stark differences become apparent in the distribution of global ancestry proportions across the region (Figure 5). Even with the lower SNP density of merged dataset 2, overall admixture patterns are replicated from the previous run and K=5 remains the best model for maximum continental contributions. As expected based on the results of PCA, ancestry proportions were most similar between Afro-Caribbean Barbadians (ACB), Haitians and LA populations who carried much higher African ancestry, and much lower Native American ancestry than Hondurans, Colombians, Dominicans and Puerto Ricans. These overall similarities and differences in ancestry proportions were found to be statistically significant at $\alpha$=0.05 for multiple comparisons between Caribbean populations (Figure 7 and Table SM6). Cubans were the only Greater Antilles population with a small East Asian component, although in much lower proportions than Trinidadians.

Higher values of K revealed structure within continental populations and by extension in the ancestral components of admixed Caribbean genomes. At K=11, the model with the lowest cv error for the second ADMIXTURE run, two components are present within African populations, differentiating the Yoruba and Esan of Nigeria from the Mende and Gambian. The East Asian Dai and Han are also separated by two major ancestry components, represented by light and dark purple in Figure 5. Among Native American groups, there is also differentiating genetic structure, most notably between the Chibchan-Paezan speakers (which includes tribes such as the Waunana, Teribe, Kogi and Embera from Central and South America) and the other populations. This within-continent structure is also visible in the admixed Caribbean genomes. Moreno-Estrada et

al. (2013) identified a Latino specific European ancestry component, exclusively shared with Iberian populations in Caribbean Latinos. In this analysis, notable structure is present between European populations, southern groups such as Italians and Iberians have larger proportions of an ancestry component (colored dark blue in Figure 5) found at lower frequencies in northern groups such as the British and the CEU population. This dark blue component is found in higher proportions among Latino Caribbean populations than in Haitians or Lesser Antilleans, consistent with previous research (Bryc, Velez, et al. 2010, Moreno-Estrada et al. 2013). At K=10 and later again at K >13, a primarily Iberian component, shared in low frequencies with other European groups but in high frequencies with Puerto Ricans is observed. This component is found in low frequencies among the LA populations. But, it must be noted that different reference European populations were used in this investigation than in the analysis conducted by (Moreno-Estrada et al. 2013). ADMIXTURE did not find unique ancestry components separating African, East Asian or South Asian ancestry contributions in the Caribbean populations, but continental ancestry contributions, especially in the case of Europe, are not homogenous across the region.

Local ancestry estimation conducted with RFMix on merged dataset 5 revealed large variability in the distribution of ancestry proportions across admixed LA genomes. Figure 8 shows karyogram plots for four individuals with varying continental ancestry proportions. Individual D has the highest proportion of Native American ancestry observed in the sample. Individual B has what appears to be recent South Asian ancestry, based on visual assessment of the large size of the green haplotype blocks. However, this

observation has not been tested by modeling of haplotype tract lengths. Additional karyogram plots are in Figure SM9.

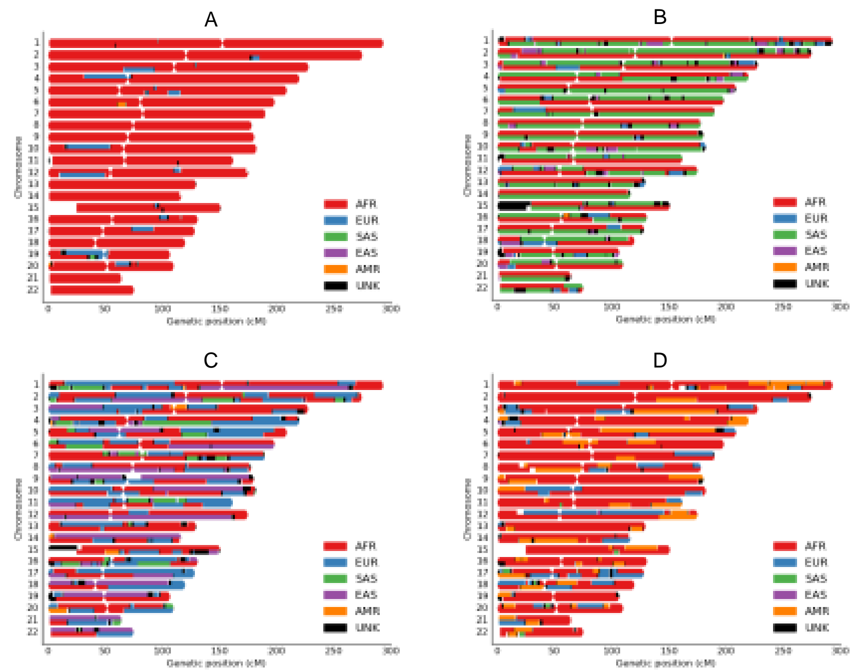


Figure 8. Local ancestry karyotypes of four Lesser Antilles Afro-Caribbean individuals. An individual with: (A) high proportions of African ancestry. (B) high proportions of South Asian ancestry. (C) high proportions of European and East Asian ancestry. (D) high proportion of Native American ancestry.

Global ancestry proportions were averaged from local ancestry estimates generated with RFMix (Table 1). Values estimated through this method differed from those generated by ADMIXTURE by approximately 1-2 percentage points. However, this difference was not found to be statistically significant (t=-4.838E-05, df=4, $P=0.999$) and the overall distribution of ancestry proportions remained constant. Figure 9 shows average autosomal global ancestry proportions for LA Afro-Caribbeans calculated from RFMix local ancestry estimates, assuming a model of K=5 continental source populations.

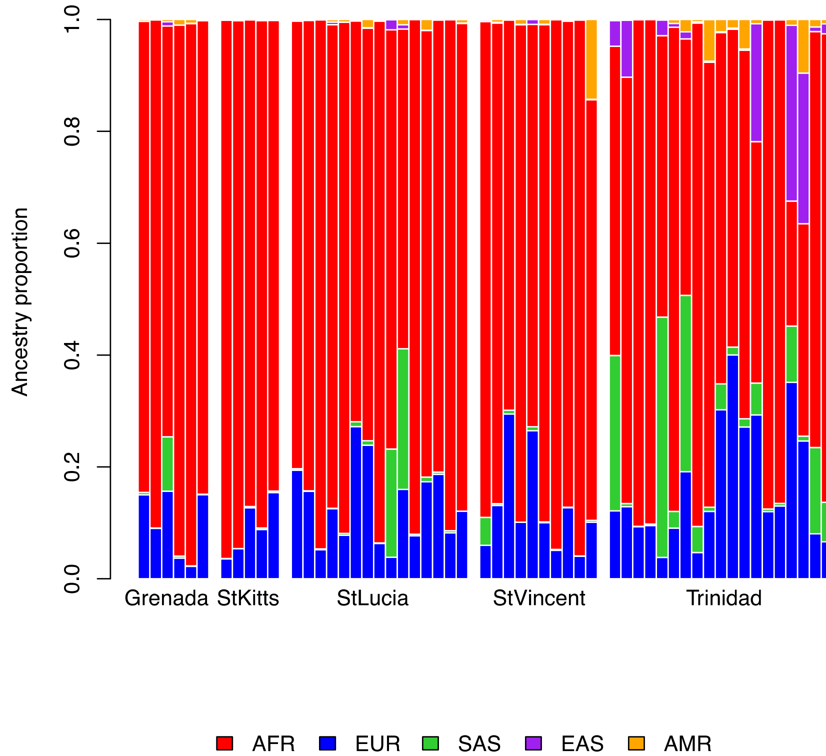


Figure 9. Global ancestry proportions were averaged per individual genome from RFMix local ancestry per-marker estimates. Local ancestry estimates were conducted with merged dataset 5 (119,277 autosomal SNPs).

2.4.2 *Sex biased admixture in the Lesser Antilles*

To identify instances of sex-biased gene flow in LA Afro-Caribbeans, global ancestry on the X chromosome was characterized separately from the autosomes. To ensure that only diploid chromosomes were compared, all males were excluded from this analysis, leaving a smaller sample size of 23 females. PCA analysis of admixed X chromosome genotypes and merged dataset 4X continental reference populations from Africa (YRI), Europe (CEU), South (ITU) and East Asia (CHB) demonstrates similar patterns as observed with autosomal SNPs. Most individuals cluster close to Africa. A few outliers, mostly from Trinidad and St. Vincent, fall closer to non-African populations

(Figure SM10). As mentioned in the Methods section, Native American X chromosomes were not available from the masked reference panel for inclusion in this, and subsequent analyses on chromosome X loci. In most cases this had the effect of inflating estimates of East Asian ancestry proportions in the admixed genotypes (Table 2).

Table 2. Ancestry proportions for chromosome X in Lesser Antilles Afro-Caribbean populations (average for females only).

| **Population** | **N** | **AFR** | **EUR** | **SAS** | **EAS** |
|---|---|---|---|---|---|
| Grenada* | 1 | 93.25% | 5.18% | 0.00% | 1.56% |
| St. Kitts* | 1 | 94.02% | 5.98% | 0.00% | 0.00% |
| St. Lucia | 11 | 92.29% | 2.34% | 3.60% | 1.77% |
| St. Vincent | 7 | 93.90% | 4.93% | 0.48% | 0.68% |
| Trinidad | 3 | 71.32% | 12.31% | 8.20% | 8.16% |
| Lesser Antilles | 23 | 90.16% | 4.71% | 2.94% | 2.19% |

* Grenada and St. Kitts only had one female each.

ADMIXTURE was used to explore the fit of admixture models for several values of K on the X chromosome, using the same four reference populations mentioned above. The lowest cv error was observed at K=3, although the ITU population cannot be differentiated from CHB and CEU at this level (Figure SM11). The ADMIXTURE program manual recommends using at least 10,000 SNPs to differentiate between continental populations, therefore it is unclear whether this is an artifact of the lower SNP density of the X chromosome dataset (6,175 SNPs after LD pruning). At higher values of K, the African component observed in LA females becomes differentiated from the reference YRI population. This pattern may be due to structure within the African ancestry of admixed LA X chromosomes, or it may be another artifact of low SNP density in this analysis.

Comparisons of autosomal versus X chromosome global ancestry proportions revealed a significant difference in African, European and South Asian ancestry proportions across the two genetic systems at α=0.05 (Table SM7). Visual comparison using bar and boxplots suggests that African and South Asian ancestry are increased in the X chromosomes of LA Afro-Caribbean females, while European ancestry is higher on the autosomes (Figure 10). These data suggest that sex-biased mating patterns, where mostly European males reproduced with African, Native and South Asian females, played a large role in shaping the genetic diversity of LA Afro-Caribbean populations. Additional boxplots comparing distribution of autosomal versus X chromosome global ancestry proportions between island populations are shown in Figure SM12.

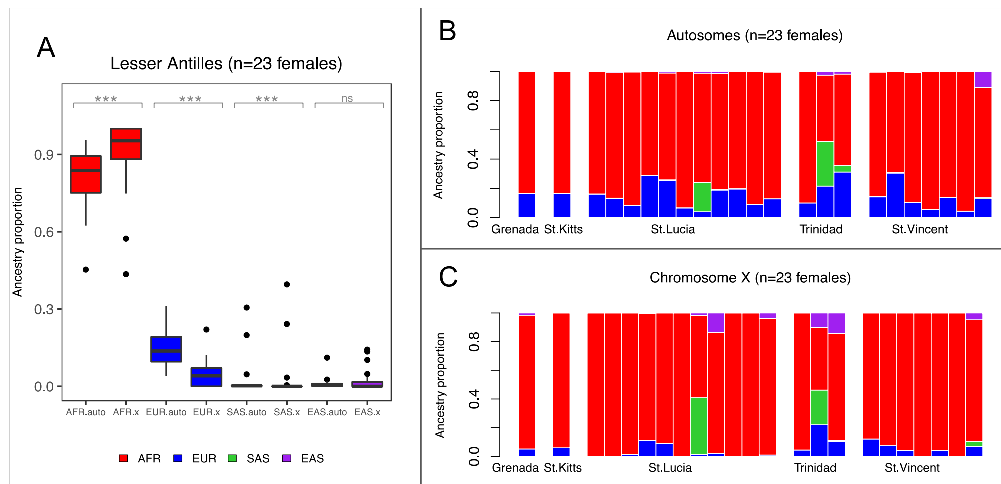


Figure 10. (A) Boxplot comparison of global ancestry proportions derived from local ancestry estimation. Significant comparisons in Wilcoxon paired test noted with asterisks: *** = $P \leq 0.001$, ** = $P \leq 0.01$, * = $P \leq 0.05$, ns = $P > 0.05$. (B) Stacked bar plot of autosomal global ancestry proportions. (C) Stacked bar plot of chromosome X ancestry proportions.

2.4.3 *Scans of natural selection: Long range haplotype methods*

Two long range haplotype-based measures, iHS and XP-EHH, were applied to identify candidate windows under positive selection in the genomes of LA Afro-Caribbean populations. The iHS test was deployed to detect signatures of incomplete selective sweeps resulting in moderate frequency alleles, and the XP-EHH between population test was used to detect signatures of complete selective sweeps where alleles are fixed in one population but remain polymorphic in the other. This difference in scope makes both tests complementary (Sabeti et al. 2007, Voight et al. 2006). The selection scans were performed on a grouped dataset of all 55 LA individuals to increase statistical power (see *Methods)*. For comparison, the tests were also performed on the four reference populations included in merged dataset 3 (YRI, CEU, ITU, CHB). Merged dataset 3 was used due to its high SNP density.

In the LA group, genome-wide scans identified thirteen 100 kb windows with high proportions of SNPs in the top 1% of |iHS| scores (Table 3). Genes within those windows were associated with cellular and organismal processes, biological regulation, and response to stimulus. Specific gene families within these candidate regions include the olfactory gene cluster on chromosome 6 and the keratin associated protein family on chromosome 17. |iHS| scores for SNP within these windows are highlighted in green in Figure 11. The PANTHER statistical overrepresentation test found that six genes associated with sensory perception of smell and with the G-protein coupled receptor signaling pathway were overrepresented in the top 1% iHS windows ($P<0.001$) (Table 4). To find signatures of selection that may be shared across populations, overlapping candidate iHS windows in the top 1% in the LA population and in the top 5% of the YRI,

CEU, ITU or CHB populations were identified as in (Pickrell et al. 2009, Cardona et al. 2014). Seven windows overlapped exclusively between LA and YRI (53% of top iHS windows in the LA group) (Table 2). Since African ancestry is the largest ancestry component in the LA population, these results are consistent with the findings of admixture and population structure analyses discussed previously. No genes were significantly overrepresented in these seven overlapping windows with respect to expected proportions in the annotated human genome.

Table 3. Top 1% 100 kb candidate regions identified in iHS analysis conducted with 55 Lesser Antilles Afro-Caribbeans.

| Chr | Window Start | Window End | SNPs in window | Proportion of SNPs with \|iHs\| > 2 | Genes in Window | YRI top 5% iHS windows |
|---|---|---|---|---|---|---|
| 1 | 63100001 | 63200001 | 50 | 0.52 | DOCK7 | |
| 2 | 17300001 | 17400001 | 55 | 0.672727 | None | |
| 2 | 21300001 | 21400001 | 74 | 0.567568 | TDRD15 | |
| 2 | 111600001 | 111700001 | 11 | 0.727273 | ACOXL | |
| 2 | 194900001 | 195000001 | 10 | 0.9 | None | X |
| 3 | 148100001 | 148200001 | 134 | 0.537313 | None | X |
| 3 | 175900001 | 176000001 | 158 | 0.506329 | None | X |
| 6 | 29000001 | 29100001 | 94 | 0.680851 | OR2W1, OR2B3 | X |
| 6 | 29100001 | 29200001 | 84 | 0.666667 | OR2J1 (pseudogene), OR2J3, OR2J2 | |
| 6 | 29200001 | 29300001 | 69 | 0.536232 | OR14J1 | |
| 8 | 9600001 | 9700001 | 26 | 0.692308 | TNKS | X |
| 17 | 39100001 | 39200001 | 38 | 0.921053 | KRT39, KRT40, KRTAP3-3, KRTAP3-2, KRTAP3-1, KRTAP1-5, KRTAP1-4, KRTAP-3, KRTAP1-1 | X |

The XP-EHH genome-wide scan identified over 500 candidate windows with high proportions of SNPs in the top 1% of scores across all pairwise comparisons: 77 in

LA vs. YRI, 108 in LA vs. CEU, 243 in LA vs. ITU, 168 in LA vs. CHB (Tables SM8 – SM11). The comparison with the lowest amount of top XP-EHH windows is LA vs YRI, probably due to the large similarity between these two populations. Most extreme XP-EHH scores were negative, indicating that the test largely detected variation that is nearly fixed in continental reference populations but remains polymorphic in admixed Lesser Antilleans (Figure 12). As with the iHS scan, most genes within the top 1% XP-EHH windows were involved in cellular or metabolic processes and responses to stimulus. But locomotion and immune system related genes were also identified in the cross-population comparisons. Known population-specific targets of selection were detected within the top candidate regions, such as the LCT gene in the LA vs. CEU comparison and the EDAR gene in the LA vs CHB comparison (Bryk et al. 2008, Bersaglieri et al. 2004). Olfactory genes were identified in all pairwise comparisons except between LA vs. YRI. This suggests that there are few differentially fixed alleles between YRI and the admixed Lesser Antilleans in the olfactory receptor gene family.

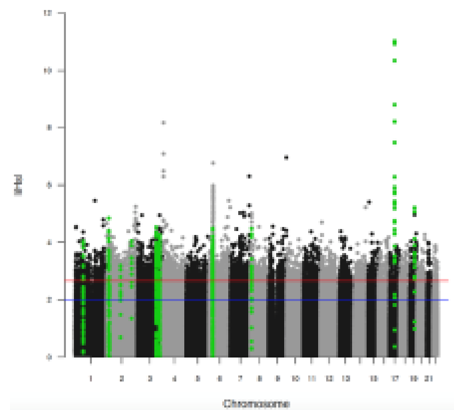

Figure 11. Manhattan plot of standardized genome-wide |iHS| scores for Lesser Antilles Afro-Caribbeans (N=55). The blue line indicates the theoretical threshold of |iHS| > 2 and the red line indicates the empirically determined top 1% quantile threshold. SNPs in green are within top 1% 100 kb candidate windows on chromosomes 1-3,6,8, 17 and 19.

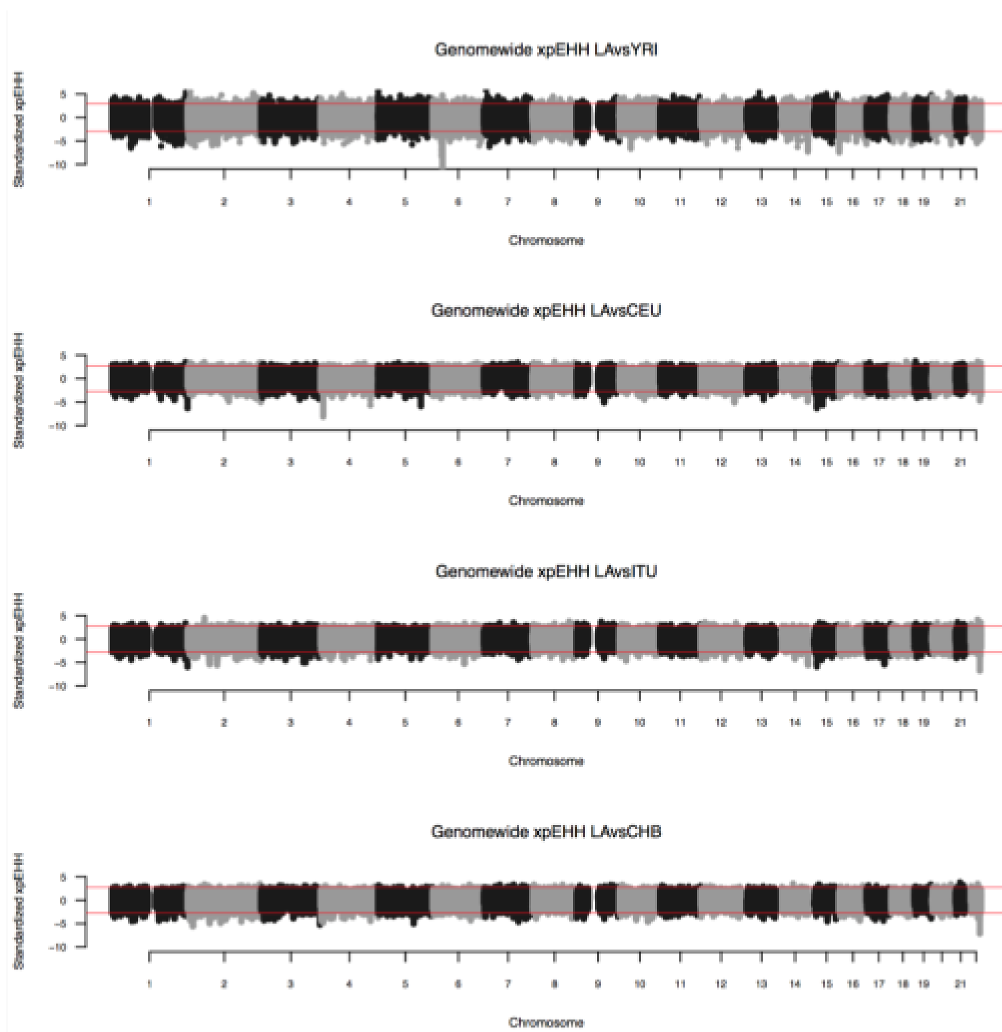


Figure 12. Manhattan plots of standardized genome-wide XP-EHH scores across four cross-population comparisons. Values above the red line were empirically determined to be in the top 1% quantiles.

The PANTHER statistical overrepresentation test identified eight genes associated with cellular defense response, antigen processing and antigen presentation that were overrepresented in the top 1% XP-EHH windows in the LA vs. YRI comparison (*P<0.001*) (Table 4). Seven of the genes belong to the HLA gene cluster on chromosome 6, and the eighth, NCR3LG1 on chromosome 11, belongs to the immunoglobulin receptor superfamily. The test also found a significant overrepresentation of genes

associated with regulation of gene expression and epigenetic processes in top candidate windows in the LA vs. ITU comparison (*P<0.001*). These genes belonged to the histone cluster 1 H2A family on chromosome 6. Only one window was identified in the top 1% of all pairwise XP-EHH comparisons. It is located on chromosome 1 between positions 416kb to 417kb, within the length of gene SCMH1. This gene is involved in cell cycle processes and transcription regulation.

2.4.4 *Scans of natural selection: Ancestry deviations*

Given that long-range haplotype methods do not allow for identification of selection that occurs in an admixed population after the admixture event(s), a genome-wide scan for ancestry deviations was conducted on the 55 LA Afro-Caribbean autosomal genotypes. This test assumes that alleles from a given ancestry component which provide a selective advantage in the admixed group will rise in frequency after the onset of admixture and therefore create genomic segments with deviations, or skews, in local ancestry proportions relative to genome-wide averages. For all 55 LA individuals 119,277 autosomal loci at the intersection of merged dataset 5 populations were scanned for ancestry deviations. Although this is a much smaller dataset than the one used for the iHS and XP-EHH tests, it has more than the minimum 3,000 markers recommended by Bhatia et al. (2014) for testing ancestry deviations in African-American populations. As recommended by Bhatia et al. (2014) genomic segments were considered candidate regions under selection if deviations in local ancestry exceeded 4 standard deviations (SD) from average global ancestry for the whole LA sample.

Table 4. Significant results of PANTHER statistical overrepresentation test on gene lists retrieved from top 1% candidate windows for iHS and XP-EHH analyses in Lesser Antilles Afro-Caribbeans.

| Test | PANTHER Biological Process (BP) | Expected no. of genes in top 1% windows | Observed no. of genes in top 1% windows | Over or under represented | Fold Enrichment | P-value | Genes associated to BP in top 1% windows |
|---|---|---|---|---|---|---|---|
| LA iHS | Sensory perception of smell | 0.37 | 6.00 | Over | 16.13 | 3.81E-04 | OR14J1, OR2B3, OR2W1, ORJ23, OR2J2, OR2J1 |
| | G-protein coupled receptor signaling pathway | 0.71 | 6.00 | Over | 8.46 | 1.46E-02 | OR14J1, OR2B3, OR2W1, ORJ23, OR2J2, OR2J1 |
| | Unclassified | 1.11 | 13 | Over | 1.17 | 0.00E+00 | multiple |
| LA vs. YRI XP-EHH | Antigen processing and presentation | 0.29 | 8 | Over | 27.23 | 1.96E-07 | HLA-DMB, HLA-DPB1, HLA-DMA, HLA-DOA, HLA-DPA1, HLA-B, NCR3LG1 |
| | Cellular defense response | 1 | 7 | Over | 6.99 | 1.73E-02 | HLA-DMB, HLA-DPB1, HLA-DMA, HLA-DOA, HLA-DPA1, NCR3LG1 |
| | Unclassified | 41.58 | 34 | Under | 0.82 | 0.00E+00 | multiple |
| LA vs. CEU XP-EHH | Unclassified | 57.63 | 52 | Under | 0.9 | 0.00E+00 | multiple |
| LA vs ITU XP-EHH | Regulation of gene expression, epigenetic | 0.52 | 5 | Over | 9.61 | 4.80E-02 | HIST1H2AG orthologs |
| | Unclassified | 88.09 | 72 | Under | 0.82 | 0.00E+00 | multiple |
| LA vs. CHB XP-EHH | Unclassified | 117.32 | 117 | Under | 1 | 0.00E+00 | multiple |

The scan identified 566 loci across seven chromosomes with ancestry skews exceeding the 4 SD significance threshold (Table 5). All significant markers had an excess of Native American (AMR) or East Asian (EAS) ancestry. However, these deviations may not necessarily be due to the action of natural selection. Many of the significant loci were found at the beginning or end of chromosomes in regions where local ancestry estimation has the most error (Figures 13 & 14) (Bhatia et al. 2014). Further, loci with significant deviations in AMR ancestry often also had significant deviations in EAS ancestry suggesting that these scores co-varied. Lastly, the ancestry skew test assumes that deviation scores and ancestry proportions approximate a normal distribution (Chimusa et al. 2015). However, probability densities of EAS and SAS ancestry proportions and deviation scores were both right skewed (Figure SM13). Thus, it seems that for a given genomic region, when proportions of an ancestry component are small, any change in allele frequency makes loci within the region more likely to cross the significance threshold, leading to the identification of false positives. This spurious signal effect was further magnified with decreasing sample size as was observed when testing for significant deviations in each of the island populations individually (Figure SM14). In the ancestry deviation test conducted with St. Kitts, the population with the smallest sample size, 6,100 significant skews across 15 chromosomes were identified and every ancestry proportion was overrepresented at least once (Table 5).

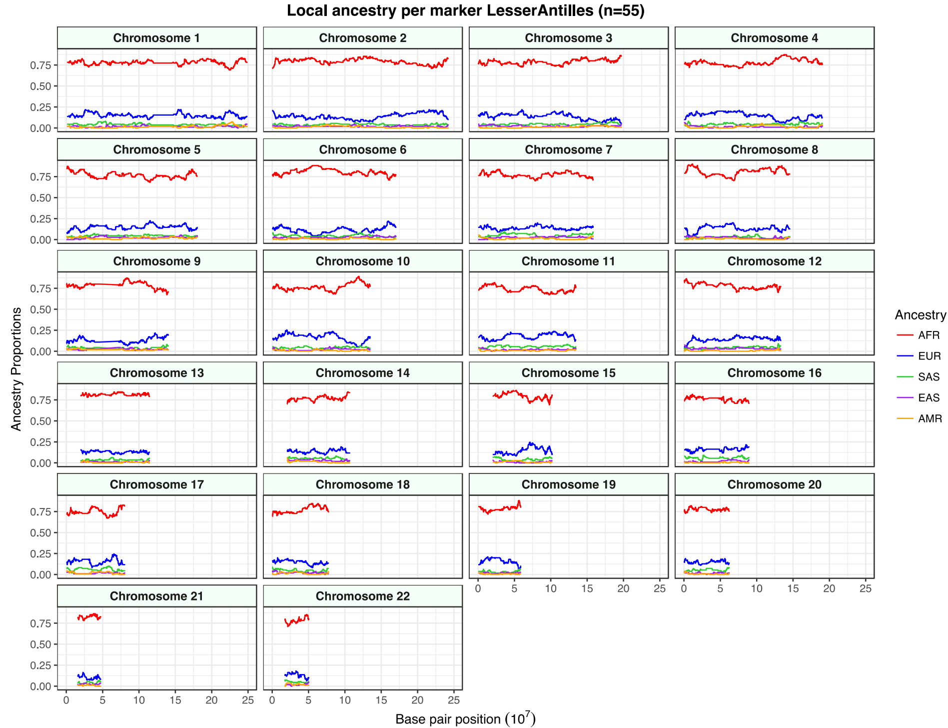


Figure 13. Genome-wide local ancestry proportions per chromosome in Lesser Antilles Afro-Caribbeans (n=55). African ancestry is >50% for all loci (119,277 autosomal SNPs).

To examine whether test results were influenced by the number of SNPs examined, the ancestry deviation scan was repeated on all 55 LA individuals with 852,887 autosomal loci at the intersection of merged dataset 4 populations. This analysis excluded Native American ancestry. In this case, no locus passed the 4 SD significance threshold. However, when the scan was repeated for individual island populations, the same situation observed above arose again. Multiple loci, usually with excess East Asian or South Asian ancestry, crossed the threshold as sample size decreased (Table 5). Therefore, I consider the results from this test to be unreliable and conclude that with these methods it is impossible to determine whether selection has occurred since admixture in the tested populations.

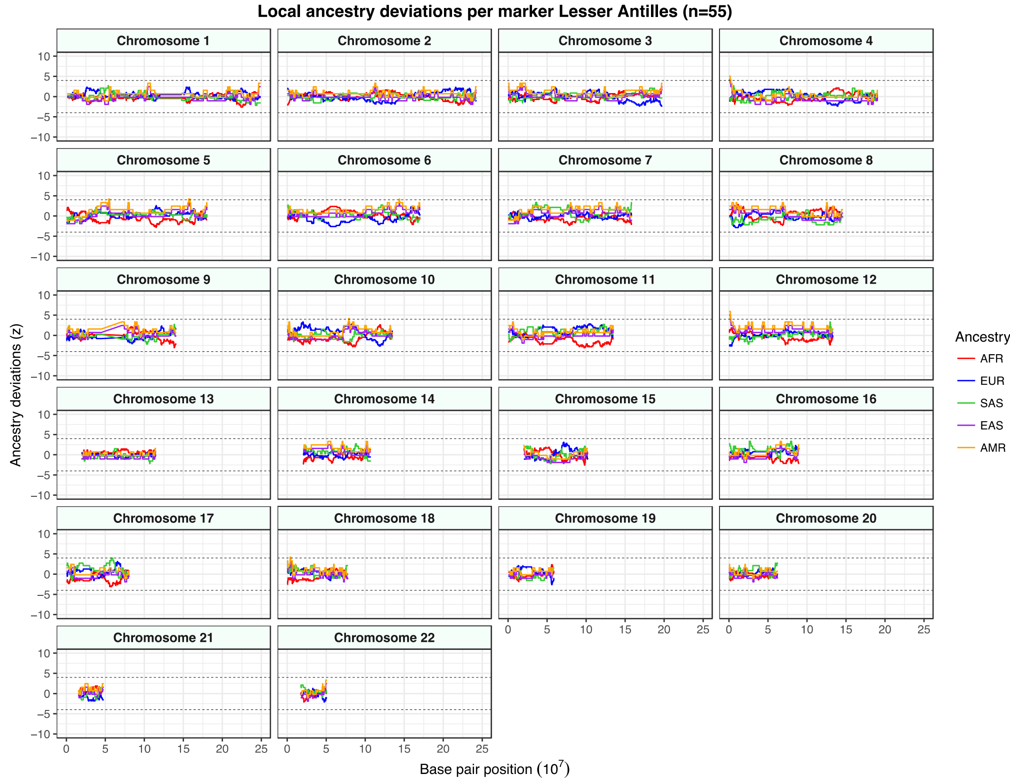


Figure 14. Standardized genome-wide local ancestry deviations in Lesser Antilles Afro-Caribbeans (n=55). The dotted lines represent 4 standard deviations from average global ancestry proportions (119,277 autosomal SNPs).

Table 5. Autosomal markers with significant ancestry skews.

| Dataset | Population | N | Markers with significant ancestry deviations | Ancestries with over or under representation |
|---|---|---|---|---|
| Merged Dataset 5 (119,277 SNPs) | St. Kitts | 5 | 4,557 | AFR, EUR, SAS, EAS AMR |
| | Grenada | 6 | 6,100 | SAS, EAS AMR |
| | St. Lucia | 15 | 1,127 | SAS, EAS AMR |
| | St. Vincent | 10 | 7,233 | SAS, EAS AMR |
| | Trinidad | 19 | 2,328 | SAS, EAS AMR |
| | **Lesser Antilles** | **55** | **566** | **EAS, AMR** |
| Merged Dataset 4 (852,887 SNPs) | St. Kitts | 5 | 24,016 | AFR, EUR, SAS, EAS |
| | Grenada | 6 | 3,808 | EAS, SAS |
| | St. Lucia | 15 | 14,316 | EAS, SAS |
| | St. Vincent | 10 | 11,190 | EAS, SAS |
| | Trinidad | 19 | 254 | SAS |
| | **Lesser Antilles** | **55** | **0** | **NONE** |

## 2.5 Discussion

### 2.5.1 *Population structure and genetic admixture in Lesser Antillean Afro-Caribbeans*

Results from multivariate and admixture deconvolution analyses are consistent with previous genetic research, which has found that post-contact migrations had a large influence on the genetic structure of contemporary LA Afro-Caribbean populations (Benn-Torres, Kittles, and Stone 2007, Benn-Torres, Stone, and Kittles 2013). All the sampled groups had high proportions of African ancestry in autosomal loci, but patterns varied between islands. While St. Kitts Afro-Caribbeans exhibited a pattern of two-way admixture with African and European components, Trinidadian Afro-Caribbeans represent the other extreme of the spectrum, with the lowest proportion of African ancestry and larger proportions of other ancestry components. Although some of these differences may be driven by the larger sample size of Trinidadians in this dataset, the results are consistent with the historical record, and with previous genetics research, and likely reflect the effect of heterogeneous population histories across the Antilles.

Estimates of the African slave trade to the British Caribbean colonies indicate that approximately 2 million slaves were brought to the islands between the 15$^{th}$ and 19$^{th}$ century (Pérotin-Dumon 1999, Sheridan 1972). However, the volume of the trade to individual colonies often varied depending on when each island was settled and its role within the plantation monoculture export system. For instance, St. Kitts and Barbados were settled in the early 1620s as tobacco and sugar exporting colonies. Both islands became intensive plantation societies early in the colonial period and had higher rates of slave importation than the other Lesser Antillean islands (mostly due to high slave mortality rates) (Rogozinski 2008, Walvin 2013, Pérotin-Dumon 1999). In contrast, St.

Vincent was not fully settled by Europeans until the 16th century due to strong resistance from resident indigenous populations. Trinidad was a Spanish colony until 1763, when the Treaty of Paris ceded the island, along with St. Vincent, Grenada and Dominica, to the British. Before this time Trinidad did not have a large planation economy and still had a substantial indigenous population (Rogozinski 2008, Morgan 2007). Thus, planation monocultures and intensive slave importation arrived at different times to each of these islands, which could explain some of the variation in African ancestry proportions observed today. Lastly, it must be emphasized that the present sampling was skewed towards individuals who self identify as Afro-Caribbean and therefore African ancestry proportions may be higher than in the general population.

Low proportions of Native American ancestry were observed in all LA Afro-Caribbean groups. This pattern may be due to the strong reduction in population size and forced re-locations of LA native groups during the early contact period (Anderson-Córdova 2005, Rivera-Pagán 2003). Proportions of Native American ancestry in St. Vincent and Trinidad, where indigenous populations still live today, are slightly higher than in the other groups, but this difference was not statistically significant ($P>0.05$). These results are largely consistent with those obtained previously with uniparental loci and ancestry informative markers that reported a tri-hybrid pattern of genetic structure for Afro-Caribbeans in the LA (Benn-Torres, Kittles, and Stone 2007, Benn-Torres, Stone, and Kittles 2013). However, estimates of Native American ancestry reported here for all populations are much lower than those listed by Benn-Torres, Stone, and Kittles (2013). This may stem in part from the lower sample sizes used in this study or from conflation

of East Asian and Native American ancestry components for some populations in estimates conducted in the present research.

Varying proportions of South and East Asian ancestry were found in admixed Caribbean populations. The highest proportions were identified among Trinidadian Afro-Caribbeans but other LA populations, as well as some Cubans and at least one Afro-Caribbean Barbadian also carried low proportions of South and East Asian ancestry. This is consistent with historic evidence for the importation of Chinese and Indian laborers during the post-emancipation indenture ship period in the Caribbean (1834-1917) (Lee-Loy 2004, Look Lai 2004, Rogozinski 2008). East Asian immigration to the LA began in 1806 with the importation of a small group of Chinese laborers to Trinidad. It intensified after the 1850s with concerted efforts on part of the British government to recruit laborers from the Chinese provinces of Guangdong and Fujian to work in the colonial sugar industry (Look Lai 2004). Between 1853 and 1884, at least 17, 000 laborers arrived from China to the British Caribbean colonies (Look Lai 2004). Most were sent to Trinidad and British Guiana. Many also arrived in Cuba during this time (Lopez 2004). The Chinese continued migrating into the Caribbean well into the 20$^{th}$ century as communities of ex-indentured laborers prospered in the shop keeping and service industries (Laurence 2011, Richardson 2004).

Indian indentured laborers were first transported to the Caribbean in 1838 from provinces in northwest India such as Uttar Pradesh (mostly Calcutta), Bengal and Bihar, but also from southern areas such as Madras and Odisha (Samaroo 2011). More than 400,000 laborers arrived between 1838 and 1917. As with the Chinese, most worked in the sugar plantations of Trinidad, Guiana and Jamaica but some also went to other LA

British possessions (Laurence 2011, Mohammed 2002). Chinese and Indian migrations to the Caribbean were biased towards male laborers (Laurence 2011), yet in the case of Indians the government attempted to bring more women by establishing sex quotas in 1856 (Mohammed 2002). Although few Chinese women came to the Caribbean overall, historical documents indicate they migrated in larger numbers to the LA than to the Greater Antilles (Laurence 2011).

Several historical sources indicate that inter-marriage or concubinage between Indians and other ethnic groups were rare throughout the 19th century due to between-group tensions, structural barriers to cultural exchange and adherence to the caste system on the part of Hindu Indians (Mohammed 2002, Diptee 2000, Samaroo 2011). However, after the 1920s and until the present, unions between Indian men, Afro-Caribbeans, Chinese and other creoles became more common (Mohammed 2002, Birth 1997). Asian ancestry had not been previously observed in genetic studies with LA Afro-Caribbeans (Benn-Torres, Kittles, and Stone 2007, Benn-Torres, Stone, and Kittles 2013, Benn-Torres et al. 2012), but mitochondrial lineages common in northern India have been identified in Trinidadian indigenous communities (Benn-Torres et al. 2015). It is worth noting that other world regions including Portugal, Africa and the Yucatan peninsula contributed laborers (willing and otherwise) to the Caribbean islands during the 19th century (Richardson 2004, Samaroo 2011). However, at the present scale of analysis, disentangling genetic components these groups may have contributed to LA Afro-Caribbean populations from more ancient African, European or Native American ancestry is not feasible.

In addition to highlighting differences in the admixture process between LA Afro-Caribbean groups, these findings also reveal consistent differences between LA islanders and other admixed island and mainland Caribbean populations. In contrast to the pattern observed in the LA, admixed Caribbean Latinos have a three-way admixture pattern, with large to moderate contributions from European and Native American sources (Bryc, Velez, et al. 2010, Gravel et al. 2013, Homburger et al. 2015, Moreno-Estrada et al. 2014, Moreno-Estrada et al. 2013). Other mainland Latino populations such as Mexicans, Peruvians, Argentinians and Chileans have much higher proportions of Native American and European ancestry (Moreno-Estrada et al. 2014, Homburger et al. 2015). Although Caribbean Latinos carry higher African ancestry proportions than their mainland counterparts (Bryc, Velez, et al. 2010, Bryc et al. 2015), African ancestry is much higher in LA Afro-Caribbeans. Of all the Greater Antillean populations tested in this research, only Haitians were observed to carry similar proportions of African and European ancestry as the LA populations ($P>0.05$). Dissimilarities were also observed in the structure of continental ancestry components across the Caribbean. An Iberian-specific European component previously identified among admixed Latinos (Moreno-Estrada et al. 2013) was not observed in the LA Afro-Caribbeans. Sex biased ancestry patterns however are largely similar across all admixed American populations examined to date. Higher proportions of European ancestry in autosome versus X chromosome loci, and concomitant increases in Native American or African ancestry on the X are observed across admixed populations from the Caribbean, mainland Latin America and the United States (Bryc et al. 2015, Bryc, Velez, et al. 2010, Homburger et al. 2015, Moreno-Estrada et al. 2013).

Differences in ancestry patterns between Caribbean populations are consistent with the dissimilar approaches to colonization and economic development employed by European powers across the Caribbean basin. The Spaniards settled the Greater Antilles in the late 15th and early 16th centuries as gold production centers for the benefit of the Crown. Once this economic system collapsed, due to the exhaustion of mineral resources and the decline of the indigenous workforce, many colonists abandoned the islands for wealthier mainland colonies like Mexico and Peru (Moya Pons 1999). Although a few sugar and coffee plantations were established, most colonists remaining on Cuba, Santo Domingo (now Dominican Republic), Puerto Rico, Jamaica and Trinidad (then Spanish possessions) turned to cattle and subsistence farming (Moya Pons 1999, Pérotin-Dumon 1999, Heuman 1997). Therefore, the slave trade to the Spanish colonies dwindled until the 19th century sugar boom increased slave importation, especially in Cuba (Clarke 2011). On Puerto Rico, however, slave importation never reached the high levels observed in the slave societies of Saint-Domingue or Barbados (Knight 1997a).

The anthropologist Harry Hoetnik suggested that the lagging economic prosperity of the Spanish Greater Antilles during the 18th and 19th centuries relaxed social barriers between European, Native American and African settlers (especially in rural areas) and allowed for high rates of cross-cultural interaction during this time (Hoetnik 1985). European migration to the Spanish islands further increased in the 19th century with the loss of other Spanish mainland colonies and the establishment of measures (such as the Royal Decree of Graces of 1815) which incentivized relocation to Cuba and Puerto Rico (Laurence 2011, Rogozinski 2008). This situation fostered the rise of an admixed creole population with larger proportions of European ancestry compared to those observed in

Saint-Domingue and other intensive plantation societies (Clarke 2011). Later, Asian indentured servitude became a large part of the 19th century sugar boom in the British colonies and in Cuba, but Chinese laborers were not imported in high numbers in the other Spanish Antilles (Clarke 2011, Lopez 2004, Laurence 2011).

Colonization in the Lesser Antilles by the Dutch, French and British contrasted sharply with that of the Spanish colonies. Many of the islands were settled by corsairs and private trading ventures such as the West India Company. Monoculture export plantations were established in the mid-1600s, first for tobacco and then for sugar (Pérotin-Dumon 1999, Rogozinski 2008). This created high demand for workers and incentivized the importation of indentured servants and slaves in high numbers (Engerman and Higman 1997). In this aspect colonization of the LA was more similar to that of French Saint-Domingue (now Haiti), where demographic proportions were highly skewed (Knight 1997a). During the late 17th and 18th centuries, Europeans represented a small minority of the population. Most inhabitants were African and Afro-Caribbean slaves and laborers. Although creole populations did arise in these islands, historical census records indicate they were much smaller in British plantation colonies such as St. Kitts and Barbados than in the Spanish and French controlled Greater Antilles (Engerman and Higman 1997). In his treatise about the differences in 'race' and 'color' across the Caribbean, Hoetnik argues that the establishment of strict racial boundaries between groups in the non-Spanish islands was necessary for the maintenance of the plantation system and therefore disencentivized inter-mixing and cross-cultural interaction (Hoetnik 1985).

This discussion does not mean to imply that African populations did not significantly contribute to the biological and cultural development of present day Caribbean Latinos or that the slave system in these islands was milder or any less brutal to its victims, there is ample evidence for the contrary as seen in (Scarano 1992, Andújar 2006, Nistal-Moret 1984, Moreno Fraginals 1977). African ancestry proportions in Greater Antilles Latinos represent a large part of extant genetic variation and have even been shown to vary by geographic location and socio-economic status (Via et al. 2011, Moreno-Estrada et al. 2013, Marcheco-Teruel et al. 2014, Martínez-Cruzado et al. 2005, Mendizabal et al. 2008). But still, the large differences in proportions of continental admixture and patterns of genetic structure between the Greater and Lesser Antilles populations included in this research point towards a large effect of divergent post-contact histories on genetic diversity across the Caribbean basin.

Lastly, it is important to note that the sampling conducted in this research was geared toward individuals who self identify as Afro-Caribbean and therefore may bias results towards increased proportions of African ancestry in the studied populations. However, as discussed above, I have observed large contributions from non-African groups to these populations. As previous genetic and anthropological research in the region has shown, ethnic-cultural identity is not necessarily predictive of genetic ancestry in the Caribbean, even though the two remain intertwined (Benn-Torres 2014, Etkins 2016, Haslip-Viera 2006, 2001).

2.5.2. *Signatures of selection in Lesser Antillean Afro-Caribbeans*

When applied to LA Afro-Caribbeans, the iHS within-population test identified 100kb selection candidate windows across seven chromosomes. Three adjacent windows with high proportions of top 1% scores were identified on chromosome 6. All six genes within these windows (five functional genes and one pseudogene) mapped to three subfamilies within the human olfactory receptor gene family: B, J and W. This class of genes was the only biological process overrepresented in the LA sample. Additionally, a single window on chromosome 17 had the highest proportion of extreme iHS scores (>90% of SNPs had |iHS|>2). This region harbors seven keratin-associated protein genes and two keratin genes.

The olfactory gene family (OR) codes for cellular receptors that interact with odorant molecules, via the G-protein mediated transduction pathway, to trigger the sensory perception of smell (Purves et al. 2001). It is the largest gene family in the human genome, related genes can be found in clusters across all chromosomes except 20 and Y (Glusman et al. 2001). OR gene subfamilies are usually found on the same chromosome, have related sequences and mediate olfaction of structurally related odorant molecules (Malnic, Godfrey, and Buck 2004). In addition, human OR genes can be actively expressed in gut tissue, germinal cells, lungs and other non-olfactory organs (Ferrer et al. 2016, Braun et al. 2007). Although they have a high rate of pseudogenization, functional genes within the OR family have been found to be under positive selection in humans (Moreno-Estrada et al. 2008, Gilad et al. 2003, Nielsen et al. 2005), especially in Africans and African-Americans, who carry a larger repertoire of OR alleles than other world populations (Williamson et al. 2007, Gilad and Lancet 2003). All six olfactory genes

found within the top 1% iHS candidate regions in LA Afro-Caribbeans have been previously identified as belonging to the MHC-linked OR cluster (Ehlers et al. 2001). Alleles at these genes are highly polymorphic in humans, possibly due to their role in distinguishing multiple odor types (Gilad and Lancet 2003). All five of the functional genes may also be involved in mediating MHC-related mate recognition preferences (Ziegler 1997, Younger et al. 2001, Ehlers et al. 2001). One gene, OR2J3, has specifically been associated with detection of *cis*-3-hexen-1-ol odor compounds ("grassy" smells) (McRae et al. 2012).

Keratin (KRT) and keratin associated protein (KRTAP) genes are involved in the production of filaments that build epithelial cells and hair fibers, respectively (Bragulla and Homberger 2009, Moll, Divo, and Langbein 2008). KRTAP genes are only found in mammals. Humans, despite being largely hairless, have retained a similar number of these genes as other primates (Wu, Irwin, and Zhang 2008). Selection at KRTAP genes has been reported across mammal species in response to ecological pressures on hair diversification (Khan et al. 2014). In humans, signals of recent adaptation have been identified in KRTAP genes among European-American populations (Nielsen et al. 2005, Williamson et al. 2007).

The cross-population XP-EHH test identified over 500 top 1% candidate windows across four pairwise comparisons between LA Afro-Caribbeans and African, European, South and East Asian reference populations. Most of these windows had high proportions of SNPs with negative XP-EHH values, which suggests the test largely detected genomic regions under positive selection in the reference populations. Only one 100kb window, harboring gene SCMH1 on chromosome 1, overlapped across all comparisons. SCMH1

is part of the polycomb group of regulatory genes which influence development by silencing homeotic HOX genes that control the timing and pattern of structures in embryonic development (Berger et al. 1999). As a regulatory gene, SCMH1 is expressed in the testes during meiosis and influences the formation of spermatocytes (Khan et al. 2014). This gene is highly conserved across multiple organisms including flies and humans (Levine et al. 2002). Given its importance in development, SCMH1 may represents a target of selection across multiple, if not all, human populations as well as in many other organisms, but this has not been formally tested.

Two biological processes were found to be significantly overrepresented in the XP-EHH tests. When comparing the LA Afro-Caribbeans to the Yoruba, immune related loci associated with antigen presentation and cellular defense response were overrepresented. This included six HLA genes on chromosome 6 and the natural killer cell cytotoxicity gene (NCR3LG1) on chromosome 1. This is consistent with previous research which has found that African populations have the highest diversity of HLA haplotypes (Gourraud et al. 2014, Prugnolle et al. 2005). It is likely that HLA variants under strong selective constraints in Africa have risen to high frequency in the Yoruba but remain polymorphic in the LA Afro-Caribbeans. Separately, five HIST1H2AG orthologs on chromosome 6 associated with epigenetic processes and gene expression were overrepresented in the comparison between LA Afro-Caribbeans and the Tamil.

Haplotype homozygosity patterns detected by tests such as iHS and XP-EHH can be strongly affected by demographic history, population differences in recombination rates and linkage disequilibrium patterns in addition to selection (Sabeti et al. 2007, Wegmann et al. 2011, Pemberton et al. 2012). In LA Afro-Caribbeans, between

population overlap in selection signals for both tests is consistent with the patterns detected in the global ancestry and admixture analyses. The Yoruba are the only population that shares iHS candidate windows with LA Afro-Caribbeans. The comparison between these two populations also had the lowest amount of top XP-EHH windows, which means that selection signatures are most similar between these two groups than between the LA individuals and the other three populations. These results suggest there is a strong effect of population history, given the high degree of African ancestry present in the Afro-Caribbean individuals. Candidate selected variants and associated haplotypes may have been under selection in ancestral populations in Sub-Saharan Africa and have then been passed on, through gene flow, to the admixed population. Whether sufficient time has passed since the onset of admixture for haplotype decay to occur in LA Afro-Caribbeans remains to be explored. Admixture events between African and Caribbean Latino populations have been dated to between 15 and 7 generations ago (Moreno-Estrada et al. 2013). Future efforts to date admixture pulses or identify instances of continuous admixture in LA Afro-Caribbeans may provide more information on expected homozygosity patterns over time given recombination rates in these populations.

In addition to using long range haplotype methods, this research employed ancestry deviation scans to identify signatures of selection that could have arisen since the onset of admixture between continental populations. This approach has been applied to identify targets of recent local adaptation in several admixed human populations including Latin Americans (Deng et al. 2016, Rishishwar et al. 2015, Tang et al. 2007, Zhou, Zhao, and Guan 2016, Brisbin et al. 2012), African Americans (Bhatia et al. 2014,

Jin et al. 2012) and South Africans (Chimusa et al. 2015, Daya et al. 2014), as well in animal species such as cattle (Bahbahani et al. 2015, Gautier and Naves 2011). In the Americas, ancestry deviation scans have identified anywhere from one (Zhou, Zhao, and Guan 2016), to three (Deng et al. 2016, Tang et al. 2007) to 6,372 (Rishishwar et al. 2015) genomic regions carrying putative signals of positive selection since European contact. In most cases, candidate loci are involved in immune related processes. For instance, a signature of excess African ancestry at the HLA locus on chromosome 6 has been identified across multiple studies conducted with admixed Puerto Ricans, Colombians, Mexicans, and Ecuadorian, as well as in US African Americans (Rishishwar et al. 2015, Brisbin et al. 2012, Tang et al. 2007, Zhou, Zhao, and Guan 2016). But ancestry skews in Latin American populations have also been identified in non-immune related genes that are known targets of selection in continental populations, such as *EDAR* and *MAPK10* among East Asians (Rishishwar et al. 2015).

In the LA Afro-Caribbeans, the results of the ancestry deviation tests were inconclusive. As discussed in the *Results* section, the first round of analysis, which included five ancestral populations, identified 566 autosomal loci with excess Native American and East Asian ancestry. However, the second round of analysis, which excluded Native American populations, failed to replicate this pattern and did not find any evidence of selection after admixture. These results are consistent with reports by (Bhatia et al. 2014, Price et al. 2008) who were not able to replicate previous findings of ancestry skews among African Americans and Puerto Ricans. These authors have cautioned that spurious signals can arise in ancestry deviation tests due to improper accounting for long range LD, failure to consider the effect of genetic drift, use of lenient

significance thresholds or errors in local ancestry estimation (Bhatia et al. 2014, Price et al. 2008). This last point is particularly problematic in the Caribbean because available Native American reference panels, which are composed of extant indigenous groups from the mainland Americas, may fail to adequately represent the genetic diversity of pre-contact island populations who are thought to be the ancestors of modern admixed groups (Pasaniuc et al. 2013). To address this problem, Zhou, Zhao, and Guan (2016) developed a method which reconstructs Native American haplotypes from fragmented haplotype blocks present in descendant admixed populations. However, this methodology requires a large proportion of Native American ancestry to be present in the target, admixed population and therefore could not be used in the present study.

Thus, it is currently unclear if selection has acted since the onset of admixture in LA Afro-Caribbean populations. Future efforts to test whether selective processes have affected current ancestry proportions in the LA would likely benefit from redoing the analyses after removing all sites in LD across ancestral populations and from usage of a Native American reference panel that incorporated genetic data from pre-contact Antillean populations. In addition, simulating selection under complex admixture scenarios, perhaps using methods similar to those implemented by the program Admixem (Cui, Schumer, and Rosenthal 2016), would likely allow for more robust testing (Bhatia et al. 2014). Lastly, an underlying caveat of the implementation of long range haplotype and ancestry deviation selection tests in this investigation is that all LA Afro-Caribbeans were grouped together and considered as one panmictic population. This grouping may obscure differences in haplotype homozygosity and deviation patterns that may have arisen between island populations. Collection of additional samples from each island in

the future may be a way to address this issue and to gain increased statistical power to re-evaluate this and other future hypotheses.

## 2.6 Conclusion

In this research, I have used dense SNP genotyping to characterize autosomal genetic diversity in five Afro-Caribbean populations from the Lesser Antilles. Coupled with existing data from continental reference populations and from admixed populations from across the Caribbean basin, this analysis identified a complex pattern of admixture in LA Afro-Caribbeans. These populations carry inputs from up to five world regions and exhibit strong signatures of sex-biased mating. Proportions of African ancestry are high, but Native American ancestry is exceedingly low. This pattern contrasts sharply with that observed in among Caribbean Latinos such as Dominicans or Puerto Ricans, but is similar to that found in Afro-Caribbeans from Barbados and Haiti. I further observe that Trinidadian Afro-Caribbeans have the highest proportion of admixture with East and South Asian populations of all Caribbean populations studied to date. These findings suggest that genomewide ancestry and structure patterns differ across Afro-Caribbeans in the Lesser Antilles, and between Greater Antillean populations.

Genomewide selection scans conducted using long range haplotype methods identified multiple genes that may be under selection in LA Afro-Caribbeans. These included genes associated with the sensory perception of smell, cytoskeletal and hair keratin formation, and cellular developmental processes, among others. However, this work failed to identify strong signatures of selection since the onset of continental admixture in these populations. Thus, at present, it is not possible to determine whether

selection has affected the distribution of ancestry among LA Afro-Caribbeans, or whether patterns of local adaptation have arisen in the last five centuries since European contact.

Characterizing how migration and admixture shapes genetic diversity is vital for understanding human evolution, history and health. This is especially true in world regions that have undergone recent and dramatic demographic shifts, such as the Caribbean. Findings from this project underscore the large impact of post-contact migrations, driven by European colonization, the African Slave Trade and modern labor movements, on the genomic diversity of Caribbean islanders. In addition to its anthropological or historical importance, such knowledge is essential for informing the identification of medically relevant genetic variation in these populations and for driving their inclusion into personalized medicine and clinical genomics.

# CHAPTER 3: COMPARATIVE PERFORMANCE OF ANCIENT DNA EXTRACTION METHODS IN REMAINS FROM TROPICAL ENVIRONMENTS

## 3.1 Abstract

The tropics harbor a large part of the world's plant and animal diversity, and they have a long history of human habitation. But, due to the adverse effects of warm and humid conditions on ancient DNA preservation, successful DNA recovery in archaeological and historic remains from these regions is uncommon. At least some of the limitations that constrain tropical ancient DNA research may be alleviated through optimization of laboratory methods to maximize recovery of degraded genetic material. This investigation compares the performance of two methods for ancient DNA extraction on six ancient and historical teeth from tropical sites in East Africa and the Caribbean. The first, Method D, was designed for recovery of short, degraded DNA fragments from skeletal remains. The second Method H, modifies the first by adding an initial EDTA wash, and an extended digestion and decalcification step. Results indicate that both methods successfully recover ancient DNA from the tested remains, and there is no significant difference in endogenous content or damage profiles in shotgun libraries built with either extract. However, Method D samples are enriched in smaller DNA fragments and have higher GC content. Further, samples extracted with this method had a significantly higher enrichment factor after mitochondrial target capture. Since surviving DNA in ancient or historic remains from tropical contexts is likely to be extremely fragmented, these results suggest that at present, Method D is best suited for extractions with this material. However, additional optimization of extraction conditions and further

testing of Method H with different types of samples may allow for improvement of this protocol in the future.

## 3.2 Introduction

Ancient DNA (aDNA) is the genetic material that survives in ancient biological tissues such as bone, teeth and dental calculus (Warinner et al. 2014, Kaestle and Horsburgh 2002). Due to the variety of taphonomic and diagenetic processes that take place in an organism's remains after death, aDNA is inherently a low quality and low quantity source of genetic material, and is very susceptible to external contamination (Hofreiter et al. 2001, Pääbo 1989, Gilbert et al. 2006, Allentoft et al. 2012, Briggs et al. 2007, Deagle, Eveson, and Jarman 2006). Previous research has shown that once cell repair functions cease, remaining DNA decays exponentially. Therefore, most genetic information in ancient samples is found in small, degraded DNA fragments (Dabney et al. 2013, Allentoft et al. 2012). Because of this, historically most aDNA studies focused on short, but informative fragments of the genome or on multicopy loci such as mitochondrial DNA (mtDNA) (Ho and Gilbert 2010). But, recent advances in extraction and next-generation sequencing methods now allow for recovery of high resolution, complete mtDNA and autosomal genomes even from remains stretching as far back in time as the early Holocene and Middle Pleistocene (Orlando et al. 2013, Meyer et al. 2016, Meyer et al. 2014, Valdiosera et al. 2006).

Despite advances in stretching the time depth for aDNA recovery, paleogenomics research still has a constrained geographical focus. Ancient DNA preservation is negatively correlated with thermal age due to the large and accelerating effect of high

temperatures on biomolecule decay and fragmentation (Lindahl 1993, Smith et al. 2001, Adler et al. 2011, Allentoft et al. 2012, Hofreiter et al. 2015). Therefore, most aDNA studies focus on biological remains from cold and temperate world regions, which have the highest chance of DNA survival (Paijmans, Gilbert, and Hofreiter 2013, Wade 2015). Even today, attempting to retrieve aDNA from remains in tropical or warm environments is risky and technically challenging (Hofreiter et al. 2015). However, recent improvements in sampling, extraction and target enrichment methods have enabled several successful efforts at aDNA recovery from tropical sites in the Caribbean, the Yucatan peninsula and South East Asia (Damgaard et al. 2015, Gamba et al. 2014, Gutierrez-Garcia et al. 2014, Kehlmaier et al. 2017, Mendisco et al. 2015, Schroeder et al. 2015). Both today and in the past, the tropics harbor a large part of the world's biodiversity and many human settlements (Buzas, Collins, and Culver 2002, Brown 2014). Therefore, optimizing and improving methods that facilitate aDNA recovery from degraded remains found in these contexts is of great interest to archaeology, paleontology, and conservation genetics, among other fields.

This study tests the performance of two extraction methods on ancient and degraded tooth samples from tropical sites in East Africa and the Caribbean (Table 6). Specifically, I compare the method proposed by (Dabney et al. 2013), henceforth known as Method D, to a second approach which combines steps from previously published protocols (Figure 15), henceforth known as Method H (Warinner et al. 2014, Gamba et al. 2016, Dabney et al. 2013). Method D was designed to increase recovery of extremely short DNA fragments (up to 30 base pairs) in ancient bone and tooth extractions. The protocol is similar to previously developed methods that employ a 24-hour proteinase K

digestion to break up cell proteins, and use silica-based buffers and chaotropic salts, such as guanidine thiocyanate, to bind DNA fragments and remove inhibitors (Rohland and Hofreiter 2007, Höss and Pääbo 1993). However, in contrast to previous protocols, Method D utilizes silica spin columns and a guanidine hydrochloride binding buffer. The method has been successfully employed to recover aDNA from Late Pleistocene cave bear remains (Dabney et al. 2013), Middle Pleistocene hominin fossils (Meyer et al. 2016, Meyer et al. 2014), and a large variety of more recently dated human and animal remains, including at least one from a tropical context (Günther et al. 2015, Kehlmaier et al. 2017, Heintzman et al. 2015, Seguin-Orlando et al. 2014).

Table 6. Ancient and historical tooth samples included in chapter 3.

| **Sample** | **Site** | **Region, Country** | **Sample Age** | **Species** |
|---|---|---|---|---|
| PC E24 | Punta Candelero | Humacao, Puerto Rico | A.D. 400-600[1] | *Homo sapiens* |
| PC 117 | Punta Candelero | Humacao, Puerto Rico | A.D. 400-600[1] | *Homo sapiens* |
| T 251 | Tibes | Ponce, Puerto Rico | A.D. 616[2] | *Homo sapiens* |
| PI 67 | Paso del Indio | Vega Baja, Puerto Rico | A.D. 1022[2] | *Homo sapiens* |
| PI 388 | Paso del Indio | Vega Baja, Puerto Rico | A.D. 822[2] | *Homo sapiens* |
| GB 7 | Gombe National Park | Kigome, Tanzania | A.D 1966[3] | *Pan troglodytes schweinfurthii* |

[1] Approximate date, based on archaeological context.

[2] Radiocarbon date median probability calAD (Pestle and Colvard 2012)

[3] Date of individual death.

Since the publication of Method D, other modifications have also been proposed for improving endogenous DNA recovery during extraction with ancient tissues. Warinner et al. (2014) used an initial EDTA wash to remove loosely bound surface contaminants on mineralized dental calculus. They found that this wash functioned as an initial decontamination step without resulting in significant DNA loss. Other

decontamination approaches include the use of bleach and phosphate buffers in pre-digestion, although these techniques can sometimes result in DNA loss (Boessenkool et al. 2016, Korlevic et al. 2015, Ginolhac et al. 2012).

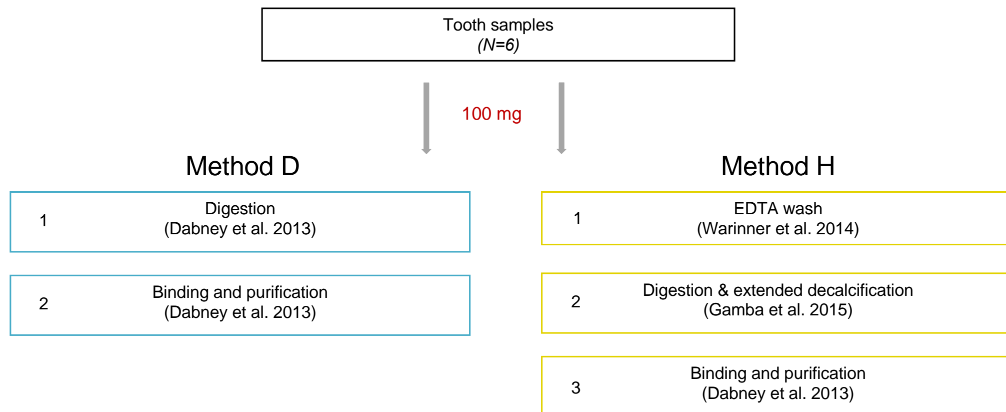


Figure 15. Schematic of extraction methods tested in chapter 3.

In a comparison between three known extraction methods, (Gamba et al. 2016) found that endogenous aDNA recovery in tooth samples extracted with Method D were similar to those obtained with a modified version of the protocol by (Yang et al. 1998), both of which use silica columns for DNA binding. The authors also found that a secondary digestion and decalcification using lysis buffer with EDTA, proteinase K and N-laurylsarcosyl detergent solution, aided in solubilizing cell proteins and resulted in increased aDNA yields. A similar finding was reported by (Damgaard et al. 2015) who observed that a brief pre-digestion (between 15 and 30 minutes) with an EDTA and proteinase K buffer was successful in reducing proportions of exogenous, contaminant DNA and enriching extracts for endogenous aDNA. Use of similar detergent solutions has been previously implemented in extraction protocols designed by (Richards, Sykes,

and Hedges 1995) and was also recently reported in extractions with petrous portion tissue (Gamba et al. 2014, Pinhasi et al. 2015).

In most of the studies referenced above, the samples tested came from temperate or cold contexts (Gamba et al. 2014, Dabney et al. 2013, Boessenkool et al. 2016); only (Damgaard et al. 2015, Pinhasi et al. 2015) included samples from tropical sites. Therefore, in this research, I build upon this body of work by focusing specifically on optimizing extraction methods with poorly preserved tooth samples from tropical contexts. I examined whether modifying Method D by adding an initial EDTA wash and an extended digestion and decalcification step (Method H) can improve endogenous aDNA recovery from samples derived from a tropical setting. To address this question, the present study examined raw DNA yields and endogenous reads recovered after shotgun sequencing from parallel extractions. Differences in base pair composition, post-mortem damage profiles and average read lengths recovered between the two methods are also characterized. Given the low endogenous content found in shotgun sequencing, I inferred that most samples examined in this work are not likely to be good candidates for whole genome enrichment. Thus, I also evaluated whether samples extracted with either method performed differently in mitochondrial enrichment capture, which is likely to have a better chance of success due to the higher proportion of mtDNA found in eukaryotic cells (Giles et al. 1980). Archaeological samples included in this research were obtained from human remains in Puerto Rico. Additionally, one historic, degraded sample from a Tanzanian chimpanzee was also included.

Study results suggest that there is no significant difference in overall aDNA or endogenous content recovered in samples extracted with either method. But libraries

sequenced from method D extracts have higher proportions of shorter DNA fragments and higher complexity. Post-mortem damage profiles are not significantly different in samples extracted by either method but average GC content was higher in samples extracted with method D. These trends were replicated in mitochondrial enriched libraries, although fragment length was skewed much higher in these data. Although both methods recovered similar amounts of endogenous DNA, since most of the archaeological samples had extremely low endogenous content (<1%), and fragment sizes averaged around 68 bp, I find that method D still provides the best chance for maximized recovery of informative ancient DNA molecules from remains buried in tropical environments.

## 3.3 Materials and Methods

### 3.2.1 *Sample and site information*

Five teeth were collected from humans interred in three open-air sites in the Caribbean island of Puerto Rico: Tibes (n=1), Paso del Indio (n=2) and Punta Candelero (n=2). All individuals date from pre-contact Ceramic Age contexts between A.D. 500-1300. Tibes and Paso del Indio are in river floodplains in the south and north of the island, respectively, and Punta Candelero is located on a coastal peninsula in eastern Puerto Rico (Curet and Stringer 2010) (Figure 16). The Köppen-Geiger system classifies Puerto Rico as a tropical monsoon environment (Kottek et al. 2006).

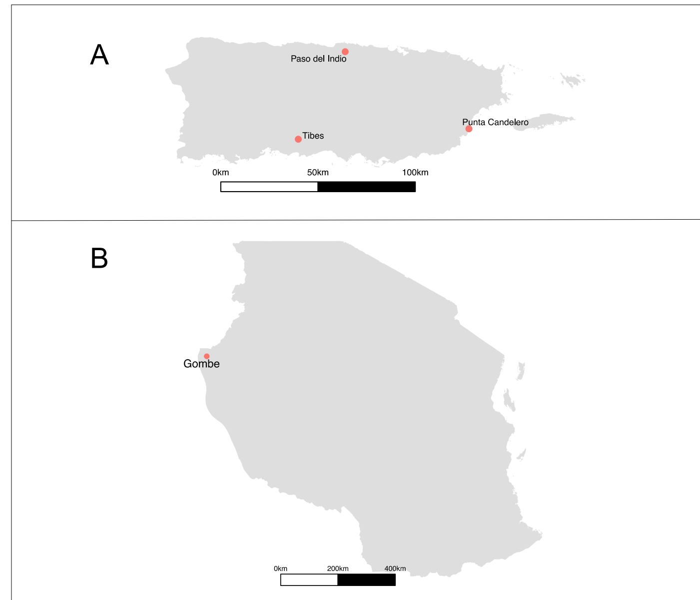


Figure 16. Sampling locations for ancient and historic tooth remains studied in chapter 3.

One tooth was collected from historic period skeletal remains of a wild *P.t.schweinfurthii* chimpanzee buried in the ground at Gombe Stream National Park in western Tanzania. The chimpanzee died of natural causes in 1966. The skeletal remains were exhumed one to two years after death and subsequently moved to the University of Minnesota. The Köppen-Geiger system classifies western Tanzania as a tropical savanna environment (Kottek et al. 2006).

3.2.2 *Sampling and DNA extraction*

Sampling and DNA extractions were conducted at the Arizona State University Ancient DNA Laboratory, a Class 10,000 clean-room facility. To eliminate surface contaminants and inhibitors, tooth samples were cleaned with a 1% sodium hypochlorite solution and the outer surface was mechanically removed with a Dremel tool (Rohland and Hofreiter 2007). Samples were also UV irradiated for 5 minutes on each side in a

UVP CL-1000 Ultraviolet Crosslinker. Teeth were sliced transversally at the cemento-enamel junction using the Dremel. The roots were covered in aluminum foil and pulverized by blunt force with a hammer as in (Schuenemann et al. 2011). Throughout the sampling process contamination controls, such as use of full body coverings, bleach decontamination and UV irradiation of tools and work area before and between uses, were implemented to minimize potential for sample contamination (Gilbert et al. 2006, Cooper and Poinar 2000).

Each sample was extracted using two different methods (Figure 15). The first method, published by (Dabney et al. 2013), was designed for recovery of short, degraded DNA fragments from skeletal remains (Method D). The second method (Method H) combines steps from existing protocols including an initial EDTA wash as in (Warinner et al. 2014), an extended digestion and decalcification step as in (Gamba et al. 2016), and binding and purification steps as in (Dabney et al. 2013) (see Appendix D for complete protocol for Method H). Approximately 100 mg of bone or tooth powder were used for each extraction. 1 µl of each extract was used to measure DNA yields in ng/µL through fluorometric quantification with the Qubit 2.0 High Sensitivity assay (Table SM4) (Simbolo et al. 2013). Extraction blanks were included throughout the process.

3.2.3 *Library preparation and sequencing*

Double stranded libraries were produced using 20 µl of each extract following the protocol published by (Meyer and Kircher 2010) with the modification that the Qiagen MinElute PCR purification kit was used in place of SPRI beads. Extraction blanks were also converted into libraries. An additional negative library control including just water

was also included to monitor contamination. 1:100 dilutions of each library were quantified using Real-Time PCR (qPCR). Reactions were run in triplicate for each library in final volumes of 20 µl with the following conditions: 10 µl of 2X Dynamo SYBR Green qPCR Master Mix with 0.3x ROX (Thermo Scientific), 1 µl of primer IS7 (5'-ACACTCTTTCCCTACACGAC-3') at 10 µM, 1 µl of primer IS8 (5'-GTGACTGGAGTTCAGACGTGT-3') at 10 µM, 7 µl of $ddH_20$, and 1 µl of the library dilution. Reactions were heated to 95°C for 10 minutes, then 40 cycles of 95°C for 15 seconds and 60°C for 1 minute for 40 cycles. A final disassociation stage was added at the end of these cycles: 95°C for 15 seconds, 60°C for 15 seconds and 95°C for 15 seconds. Quantification was performed using an ABI7900HT thermocycler and results were analyzed with SDS software. Analysis of qPCR data focused on cycle threshold values (Ct), which represent the number of qPCR cycles required for fluorescent signal to exceed background levels. Mean Ct values were averaged across all replicates per library. Non-template controls (NTC), which have no DNA, were also included in the reaction to monitor background fluorescent levels.

All libraries were double indexed and amplified for 11-20 cycles following recommendations from (Kircher, Sawyer, and Meyer 2012, Seguin-Orlando et al. 2015). To increase library complexity, four 100 µl indexing reactions were performed per library with the following conditions: 9.27 µl of 10X PCR Buffer II, 3.68 µl of 10 mM dNTPs, 2.21 µl of 10 mg/ml Bovine Serum Albumin, 9.27 µl of 25 mM Gold $MgCl_2$ solution, 2 µl of P5 indexing primer (5'-AATGATACGGCGACCACCGAGATCTACAC xxxxxxACACTCTTTCCCTACACGACGCTCTT-3') at 10000 nM, 2 µl of P7 indexing primer (5'-CAAGCAGAAGACGGCATACGAGATxxxxxxGTGACTGGAGTT

CAGACGTGT-3') at 10000 nM, 61.09 µl of $ddH_20$, 1.48 µl of AmpliTaq Gold® enzyme (Life Technologies) and 9 µl of DNA library. Reactions were heated to 95°C for 15 minutes for initial denaturation, further denaturation, annealing and elongation were performed at 95°C for 30 seconds, 58°C for 30 seconds and 72°C for 45 seconds for 9-20 cycles. Final extension was performed at 72°C for 10 minutes, and reactions were kept at 10°C. All four aliquots of each amplified library were combined and the library was purified with Qiagen MinElute PCR purification kit following manufacturer parameters with the following modification: EB buffer was preheated to 65°C, and elution was performed in 30 µl. 1 µl of each library was used for quantification with the Qubit 2.0 Broad Range assay. Purified libraries were further diluted to a factor of 1:1000 and quantified with the KAPA Library Quantification kit (Kapa Biosystems) following manufacturer instructions on the ABI7900HT thermocycler.

Fragment analysis of the amplified libraries was performed using the DNA 1000 assay on the Agilent 2100 Bioanalyzer. Heteroduplexes that arose in the libraries during indexing were eliminated through reconditioning PCR with the following conditions: 10 µl of 10X Accuprime *Pfx* reaction mix, 3 µl of IS5 primer (5'-AATGATACGGCG ACCACCGA-3') at 10 µM, 3 µl of IS6 primer (5'-CAAGCAGAAGAGGCATACGA-'3) at 10 µM, 76 µl of $ddH_20$, 1 µl of Accuprime™ *Pfx* enzyme and 7 µl of DNA library. Reactions were heated to 95°C for 2 minutes for initial denaturation, further denaturation, annealing and elongation were performed at 95°C for 15 seconds, 60°C for 30 seconds and 68°C for 1 minute for 2 cycles. Final extension was performed at 68°C for 5 minutes, and reactions were then kept at 4°C. All four aliquots of each amplified library were combined, and the library was purified again with Qiagen MinElute PCR purification kit

as detailed above. A second round of qPCR and fragment analysis was performed after reconditioning. Shotgun libraries were then sequenced on one lane of the Illumina HiSeq 2500 (2 x 100 bp reads) at the Yale Center for Genomic Analysis.

3.2.4 *Mitochondrial target enrichment capture*

In addition to evaluating overall DNA yields and recovery after shotgun sequencing, efficiency and performance of mitochondrial enrichment in DNA samples extracted with both methods were also assessed. Targeted enrichment with in-solution hybridization was performed as in (Maricic, Whitten, and Paabo 2010) with modifications as in (Ozga et al. 2016). After capture, most of the enriched libraries (all human and chimpanzee samples extracted with method H) were sequenced along with other samples across three MiSeq v2 runs (2 x 150 bp reads) at the DNASU Sequencing Core at Arizona State University. Chimpanzee enriched libraries extracted with method D were sequenced along with other samples on one lane of the Illumina HiSeq 2500 (2 x 100 bp reads) at the Yale Center for Genomic Analysis.

3.2.5 *Shotgun read mapping and processing*

Illumina sequence reads were merged and adapters trimmed using SeqPrep (https://github.com/jstjohn/SeqPrep) with a minimum overlap of 11 base pairs (bp) and a minimum length threshold of 30 bp. Read quality was assessed pre and post-merging with FastQC v.0.11.3 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). To compare sequencing results directly across extraction treatments and to control for differences in sequencer output, reads were randomly selected per sample-treatment

combination to match the lowest number of reads obtained per sample. Downsampling was performed using seqtk with default parameters (https://github.com/lh3/seqtk) (Table SM5). For human samples, the down sampled reads were mapped to the GRCh37 (hg19) reference with the mitochondria replaced by the revised Cambridge Reference Sequence (rCRS) (Andrews et al. 1999). Mapping was performed using BWA v. 0.7.5 (Li and Durbin 2009) with seeding disabled (-1 1000) and edit distance increased (-n 0.01) to improve mapping accuracy of ancient DNA reads (Schubert et al. 2014). For the chimpanzee samples, the reads were mapped to the PanTro4 assembly using the same conditions listed above. Sequence reads were filtered with SAMtools v. 0.1.19 (Li et al. 2009). Reads with mapping quality ≥ Q30 were retained, duplicates were removed with the rmdup option and reads mapping to more than one location were discarded by controlling for XA, XT and X0 tags. Damage patterns were characterized with mapDamage v.2.0.2 (Jónsson et al. 2013, Ginolhac et al. 2011). Parameters examined included deamination patterns, probability of C to T misincorporations at first position, probability of A to G misincorporations at last position, probability of a DNA fragment terminating in a single-stranded overhang ($\lambda$), probability of observing cytosine deamination in a double strand ($\delta_D$), and probability of observing cytosine deamination in a single strand context ($\delta_S$). Read quality scores were rescaled using mapDamage.

Library complexity estimates were generated using *preseq* v2.0 (Daley and Smith 2013) on downsampled bam files containing all Q30 mapped reads (including duplicates). Future experiment yield predictions were calculated extrapolating to 1e+10 total reads. For both analyses the extrapolation step size parameter was modified as: –s 100000. Summary statistics such as average %GC content, mean read depth, standard

deviation of read depth and percent of reference sequence covered were estimated on the rescaled BAM files using Qualimap v.2.2.1 (Okonechnikov, Conesa, and García-Alcalde 2016). Reads mapping to the mitochondrial genome were subset from the filtered and rescaled shotgun BAM files and re-analyzed with Qualimap as described above to obtain estimates of mitochondrial genome coverage and read depth.

3.2.6 *Mitochondrial enriched reads mapping and processing*

Sequence reads obtained after mitochondrial enrichment were downsampled in the same manner as listed above, matching the lowest number of reads obtained per sample-treatment combination (Table SM6). Read mapping and filtering followed the same parameters described above except human mitochondrial reads were mapped to the rCRS (NCBI Reference Sequence: NC_012920.1) and chimpanzee mitochondrial reads were mapped to the chimpanzee mitochondrial reference (NCBI Reference Sequence: NC_001643.1). This is the same mitochondrial sequence contained within the PanTro4 reference. Damage parameters and summary statistics were generated as listed above using mapDamage and Qualimap. To estimate the fraction of reads from the shotgun libraries that were retained after mitochondrial enrichment per each sample, the following commands were used in Bedtools as in (Ávila-Arcos et al. 2015): bedtools intersect -wa -r -f 0.9 –abam *shotgun.rescaled.bam* -b *postMT.capture.bed* > *intersected.bam*. Mitochondrial enrichment per sample was estimated as in (Cruz-Davalos et al. 2016) by calculating the ratio of the total number of unique reads mapping to the mitochondrial genome after enrichment over the total number of unique reads mapping to the mitochondrial genome in the shotgun libraries.

3.2.7 *Contamination estimates*

Nuclear coverage was too low for assessing contamination using mitochondrial data from the shotgun sequenced libraries (<3X read depth, see *Results),* so Bayesian contamination estimates per sample were performed using the mitochondrial enriched reads with contamMix (Fu et al. 2013). A set of 311 human complete mtDNA genomes were used as potential contaminant sources. As in (Fu et al. 2013), only samples with >3x read depth and >90% coverage of the mitochondrial genome were included in contamination assessment.

3.2.8 *Statistical Analyses*

Extraction yields (ng/uL), cycle threshold qPCR values, percent endogenous content (shotgun), library complexity (measured as percent distinct reads), percent GC content, average fragment lengths and damage parameters were compared for each sample across extraction treatments using paired T tests or non-parametric Wilcoxon signed rank tests. Normality assumptions were evaluated using a Shapiro-Wilk normality test (Zar 2010). However, since this test has been shown to have little power to reject the null hypothesis when sample size is low, the normality assumption was further assessed through visual examination of Quantile-Quantile plots and histograms of the difference between paired values, as recommended by (Ghasemi and Zahediasl 2012) (Figure SM15 and SM29). Correlations between variables were tested using Pearson's r as applied in the cor.test function in R. Read length distributions for both shotgun and enriched libraries were plotted as overlapping histograms.

3.2.9 *Computational resources and R packages*

This research was completed using resources from the ASU High Performance Computing Saguaro environment. All custom scripts written for this project will be available on GitHub after publication. All plots and figures were generated in R 3.2.4 using the ggplot2 (Wickham 2009), gridExtra (Auguie 2016), tidyr (Wickham 2016b) and reshape 2 (Wickham 2007) packages or with R base graphics.

## 3.4 Results

3.4.1 *DNA yields and library DNA content*

Overall DNA yields were evaluated through flourometric quantification of raw extracts (ng/uL) and through analysis of mean Ct values from qPCR of shotgun libraries (Table SM5). These analyses did not reveal significant differences in mean DNA yields or mean library DNA content between samples extracted with Method D or Method H (Table 7and Figure SM16). Low mean Ct values are indicative of high amounts of DNA template present in the library. In contrast, high mean Ct values indicate more cycles are needed to reach the threshold due to low amounts of starting DNA template. As expected, mean Ct values were inversely proportional to ng/uL quantification results for most shotgun libraries examined (Figure 17).

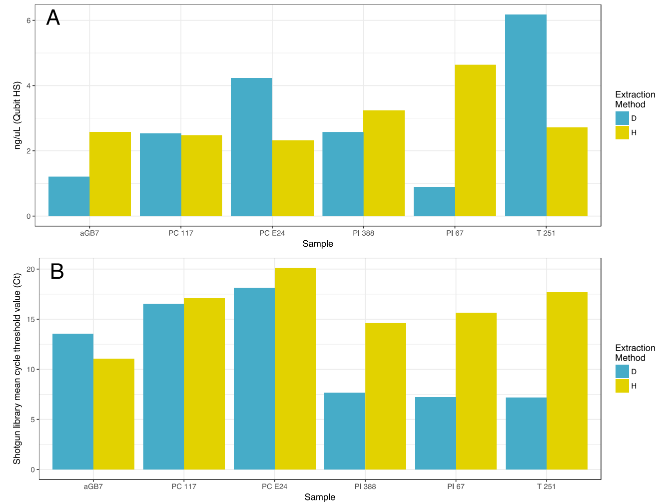


Figure 17. Per sample comparison of raw DNA yields after extraction and DNA content in shotgun libraries. (A) Double bar graph comparing extraction yields (measured as ng/uL on Qubit High Sensitivity assay) for samples extracted with both methods. (B) Double bar graph comparing mean cycle threshold values (Ct) measuring DNA content in shotgun libraries.

Table 7. Results of paired tests to determine significant differences in sample means for shotgun libraries.

| Test comparison | Shapiro Wilk test statistic | P-value | Paired test used | Test statistic | DF | P-value | P ≤ 0.05 |
|---|---|---|---|---|---|---|---|
| DNA yields (ng/uL) | W = 0.98751 | 0.9822 | T-test | t = -0.054 | 5 | 0.9589 | |
| Mean Ct value for shotgun library | W = 0.94525 | 0.7018 | T-test | t = -2.0989 | 5 | 0.0898 | |
| Percent endogenous content | W = 0.76495 | 0.0277 | Wilcoxon signed ranks | V = 6 | N/A | 0.4375 | |
| Percent distinct reads | W = 0.81734 | 0.0836 | Wilcoxon signed ranks | V = 14 | N/A | 0.5625 | |
| Average fragment length | W = 0.84833 | 0.1526 | Wilcoxon signed ranks | V = 0 | N/A | 0.0312 | * |
| Average GC content (%) | W = 0.91525 | 0.4718 | T-test | 6.006 | 5 | 0.0018 | * |
| Damage parameter DeltaD | W = 0.92583 | 0.5483 | T-test | 1.4819 | 5 | 0.1985 | |
| Damage parameter DeltaS | W = 0.98015 | 0.9523 | T-test | -0.28504 | 5 | 0.787 | |
| Damage parameter Lambda | W = 0.8583 | 0.1834 | T-test | 0.72665 | 5 | 0.500 | |

3.4.2 *Endogenous content and library complexity*

For each DNA library, between 1 and 9 million reads were obtained after shotgun sequencing. After randomly selecting the same number of reads per individual pair, between 1 and 6 million reads were analyzed per sample-treatment combination. Percent endogenous content was calculated as the proportion of unique reads (after duplicate removal and quality filtering) mapping to the reference over the total amount of down sampled reads (Table SM5). Most samples had <1% endogenous content, and similar values were obtained irrespective of extraction treatment. The exception to this pattern was sample PI-67, which had a five-fold increase in endogenous content when extracted with method H over method D (Figure 18). Both libraries from the chimpanzee sample, GB-7, had >10% endogenous content, up to sixteen-fold higher content than that found in the human libraries. These differences may be due to the younger age of the historic period chimpanzee remains compared to the archaeological human samples. Although, for both GB-7 and PI-67 the extractions conducted with Method H resulted in higher endogenous content, mean endogenous content was not found to be significantly different across the full dataset (Table 7 and Figure SM3A).

The relationship between clonality and endogenous content in the shotgun libraries is shown in Figure 19A. Sequence clonality (measured as fraction of duplicate sequence reads over total downsampled reads) is very low for all samples (<0.1%). However, replicates from PI-67 and GB-7 extracted with Method H have slightly higher clonality than Method D replicates. This suggests that libraries built with Method H may have slightly lower complexity. However, the relationship between the two variables is not linear or significant in this dataset (Pearson's $r = 0.35$, $t = 1.19$, $df = 10$, $p=0.2588$).

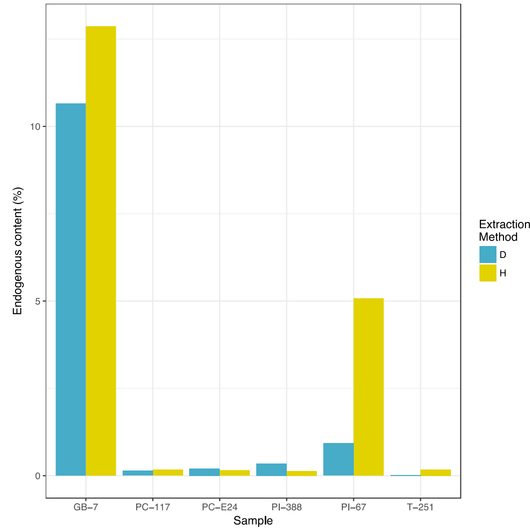


Figure 18. Per sample endogenous content in shotgun libraries. Percent endogenous content calculated as number of mapped, unique reads divided over total down sampled reads.

To examine this question further and test which method produced higher complexity libraries, I used the *c_curve* function in preseq to estimate the number of distinct reads recovered for each library. High complexity libraries have a large proportion of distinct reads that map to different parts of the reference genome. This means that there is a higher chance that more parts of the reference are covered with a single sequencing experiment. In contrast, low complexity libraries have a large proportion of distinct reads that all map to the same sites, and therefore may have a strong bias towards a given part of the genome, as well as high redundancy (Head et al. 2014). For example, libraries with high clonality levels due to excessive amplification often have many DNA fragments that align to the reference genome, but have low complexity because most of those fragments are duplicate reads. These libraries are largely uninformative because a small amount of unique DNA fragments are obtained regardless of the amount of sequencing performed (Head et al. 2014, Daley and Smith

2013). In this dataset, Method D libraries had a slightly higher mean proportion of distinct reads (96%) than Method H libraries (97%), but this difference is not statistically significant (Table 7, Figure SM17). This suggests that complexity is high overall regardless of method used. The relationship between complexity and endogenous content in the tested libraries is shown in Figure 19B. There was no significant linear association between the two values in this dataset (Pearson's $r = 0.27$, $t = 0.91$, $df = 10$, $p=0.3846$).

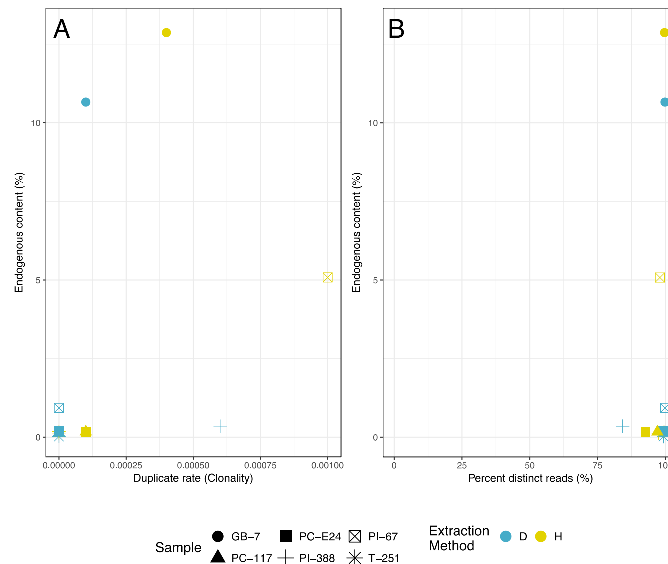


Figure 19. Scatterplots depicting relationship between endogenous content, clonality and complexity in shotgun libraries. (A) Clonality or duplicate rate as number of duplicate reads divided over total downsampled reads. (B) Complexity as the number of distinct reads before duplicate removal (measured using preseq).

In addition to estimating present complexity, the *lc_extrap* function within preseq was used to predict the expected yield for a larger sequencing effort with the same libraries. This extrapolation analysis is highly sensitive to the amount of sequence data generated, and can give false estimates with low amounts of reads (Daley and Smith

2013). Therefore, this analysis was only possible for the two samples with highest number of reads: PI-67 and GB-7. Figure 20 demonstrates that in both cases, libraries constructed with Method D extracts are predicted to yield a higher amount of complex DNA fragments with deeper sequencing (up to 10 billion reads). However, given that saturation of the complexity curve is reached early in the estimation, less than 1 million reads would be necessary to sequence all unique fragments predicted to be present in these libraries.

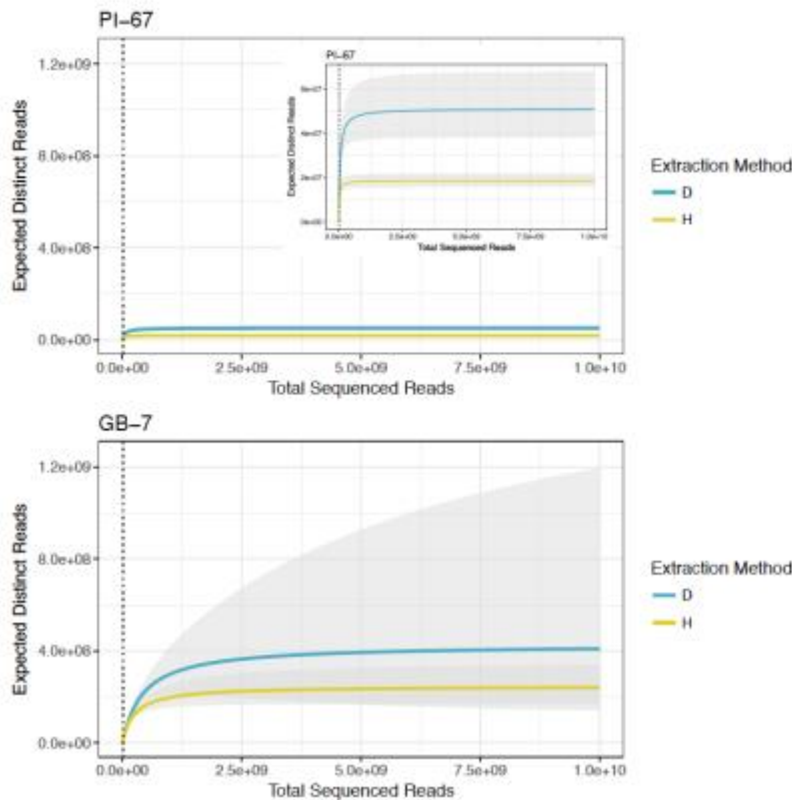


Figure 20. Extrapolation curves for shotguns library complexity estimation. Curves are shown for the two samples with highest number of reads: PI-67 and GB-7. Top inset shows zoomed in results for PI-67. Extrapolation curve and confidence interval estimation was performed in preseq using a step size of 100,000 and default sequencing effort of 10 billion reads. The dotted line denotes the number of reads randomly downsampled for each sample pair: 5.1 million reads for GB-7 and 6.8 million reads for PI-67.

3.4.3 *DNA fragment lengths and GC content*

All samples, irrespective of extraction method had average DNA fragment lengths <100 bp. This small size is consistent with expectations for degraded remains (Briggs et al. 2007, Dabney, Meyer, and Pääbo 2013) and similar to sizes obtained in previous aDNA research with tropical samples (Schroeder et al. 2015, Kehlmaier et al. 2017). However, samples extracted with Method D had smaller average DNA fragment size (68 bp) than Method H extracted samples (76 bp). Overlaid plots showing the length distribution of mapped, unique reads per sample replicate in Figure 21 demonstrate that most Method H libraries had a higher proportion of larger fragments. This pattern is replicated when comparing fragment length distributions for all reads obtained per sample, irrespective of mapping or quality (Figure SM18). No significant correlation was found between endogenous content and read length (Pearson's $r = 0.56$, $t = 2.18$, $df = 10$, $p=0.05433$) (Figure SM19).

Both mean fragment length and GC content were found to be significantly different between extraction methods (Table 7). Average GC content for all libraries was approximately 40%, however Method D libraries had higher GC content than Method H libraries by at least three percentage points: 42% versus 38%, respectively (Figure 22). A scatterplot of average DNA fragment lengths versus average %GC content clearly distinguishes between samples generated with either method. A significant negative correlation was found between average read length and GC content when considering all samples (Pearson's $r = -0.61$, $t = -2.436$, $df = 10$, $p=0.03507$); however, this was not significant when repeated separately just for samples extracted with Method D or with Method H (Figure SM20A). No significant correlation was observed between %GC and

endogenous content in the tested samples (Pearson's $r = -0.41$, $t = -1.457$, $df = 10$, $p=0.1757$) (Figure SM20B).

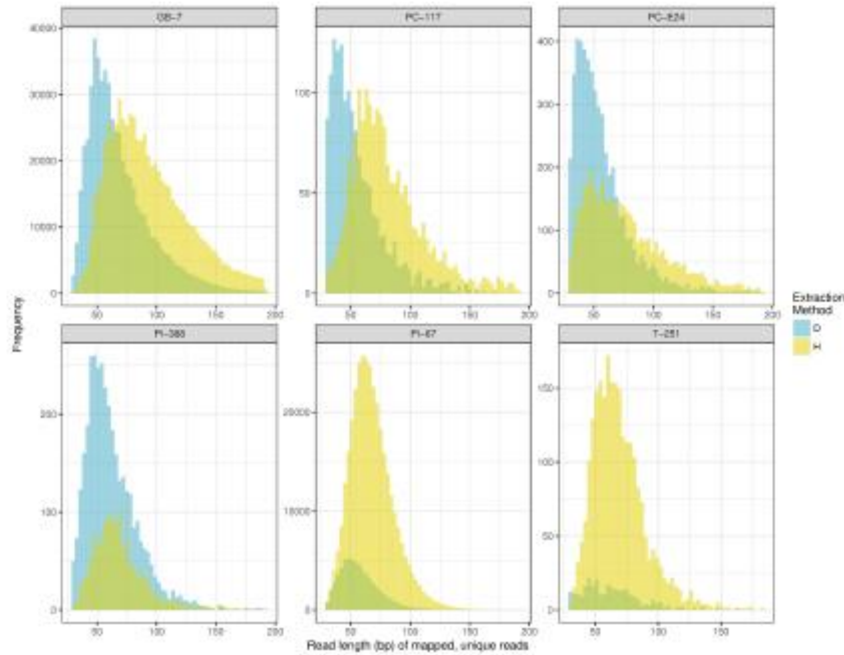


Figure 21. DNA fragment length distributions for unique, mapped reads in shotgun libraries.

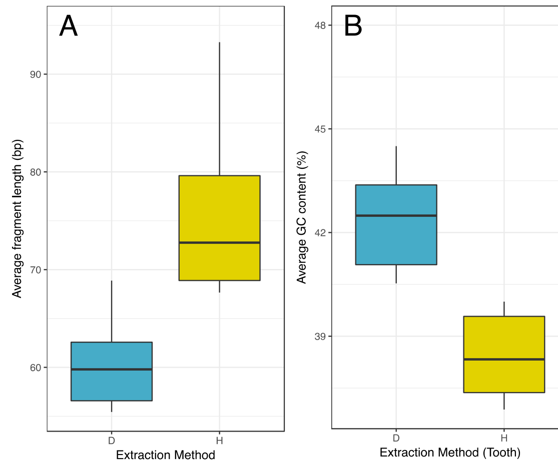


Figure 22. Boxplots comparing distributions of DNA fragment length (A) and %GC content (B) values in shotgun libraries.

3.4.4 *DNA damage parameters*

DNA damage parameters were examined to determine whether differences existed in the composition of the molecules recovered. I observed that all samples, irrespective of method, had high probabilities of C to T and A to G misincorporations caused by DNA damage at the first and last position of each fragment, respectively (>0.90) (Table SM5; see Appendix E and F for damage plots). This is consistent with known damage patterns characterizing authentic aDNA sequences (Dabney, Meyer, and Pääbo 2013, Briggs et al. 2007, Overballe-Petersen, Orlando, and Willerslev 2012). The only exception to this pattern was the PC-117 replicate extracted with Method H, which had a lower probability for both measures (~0.73). Samples extracted with Method H had lower $\delta_D$ and $\lambda$ values. Although this would initially suggest there is less probability that fragments retrieved with this method would have deamination in double strands or end in overhangs, these differences were not statistically significant (Table 7, Figure SM21). Taken together these results suggest that there is no significant distinction in the damage profiles of samples extracted with either method. Further, there was no significant correlation between average read length and any of the damage parameters tested here (Figure SM22).

3.4.5 *Mitochondrial Enrichment*

Mitochondrial enrichment per sample was calculated as the ratio of unique reads mapping to the mtDNA genome in the captured versus shotgun library (from down sampled reads). Enrichment folds ranged from 2-fold for sample PI 388, up to 638-fold for sample T 251 (both method D) (Figure SM23). For some libraries, the proportion of

unique mitochondrial reads increased by up to two-orders of magnitude after capture (Figure 23). Coverage of the mtDNA genome, mtDNA read depth and mtDNA endogenous content increased by two to three orders of magnitude for all samples except the PI 388 replicate for method H which had zero mitochondrial reads in the shotgun library. Mitochondrial enrichment factor was significantly higher in samples extracted with Method D versus H, but no significant difference was observed in mtDNA endogenous content (Table 8, Figure SM24). Between 0.1% and 12% of mitochondrial reads observed after capture were also sequenced in the shotgun libraries. In accordance with findings reported by (Ávila-Arcos et al. 2015), high mtDNA endogenous content in the shotgun libraries did not correlate linearly with high mtDNA endogenous content after capture in these samples (Pearson's $r$ =-0.201, t = -0.649, df = 10, p= 0.531) (Figure SM25).

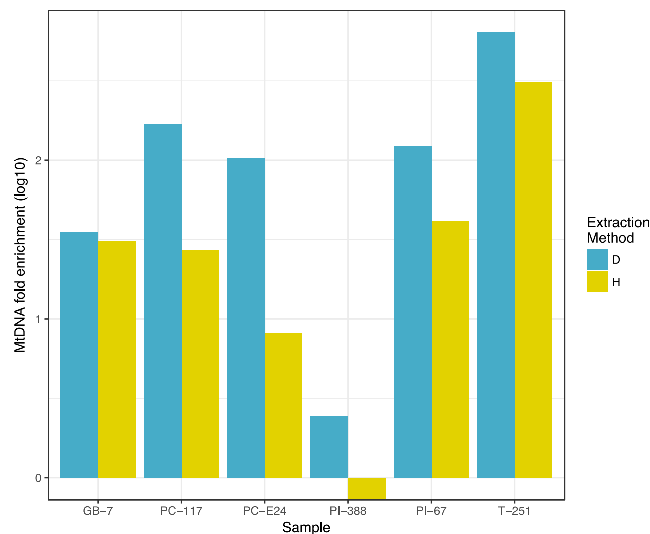


Figure 23. Per sample comparison of mtDNA enrichment. Enrichment factor shown in log10 for ease of graph interpretation. After downsampling, replicate PI-388 extracted with Method H had zero reads mapping to the mitochondrial genome.

Table 8. Results of paired tests to determine significant differences in sample means for mtDNA enriched libraries.

| Test comparison | Shapiro Wilk test statistic | P-value | Paired test used | Test statistic | DF | P-value | P ≤ 0.05 |
|---|---|---|---|---|---|---|---|
| Mitochondrial enrichment | W = 0.8568 | 0.1785 | Wilcoxon signed ranks | V = 21 | N/A | 0.0312 | * |
| MtDNA percent endogenous content | W = 0.8635 | 0.2014 | Wilcoxon signed ranks | V = 18 | N/A | 0.1563 | |
| Average mtDNA fragment length | W = 0.8089 | 0.0705 | Wilcoxon signed ranks | V = 0 | N/A | 0.0312 | * |
| Average mtDNA GC content (%) | W = 0.8243 | 0.0962 | Wilcoxon signed ranks | V = 21 | N/A | 0.0312 | * |
| Damage parameter DeltaD | W = 0.7922 | 0.0499 | Wilcoxon signed ranks | V = 6 | N/A | 0.4375 | |
| Damage parameter DeltaS | W = 0.8306 | 0.1088 | Wilcoxon signed ranks | V = 21 | N/A | 0.0312 | * |
| Damage parameter Lambda | W = 0.8488 | 0.1538 | Wilcoxon signed ranks | V = 17 | N/A | 0.2188 | |

Average mtDNA fragment length after enrichment was higher than in the shotgun libraries (Figure 24A). This is consistent with findings reported by previous studies examining efficiency and performance of whole-genome enrichment methods (Ávila-Arcos et al. 2015, Cruz-Davalos et al. 2016). There was a significant difference in fragment length in mtDNA enriched reads between extraction methods (Table 8). As with the shotgun libraries, Method H enriched samples had larger fragment sizes, on average, than Method D samples (Figures SM26-SM27). Additionally, differences in GC content post enrichment were similar to those observed in the shotgun libraries, samples extracted with Method D had significantly higher average GC content than Method H extracted samples (Figure 24B).

DNA damage patterns for post capture mtDNA reads, for the most part, mirrored those observed in the shotgun libraries (Table SM6). All human samples had high probabilities of C to T and A to G misincorporations at the first and last positions of DNA fragments (>0.90). However, for both replicates of the chimpanzee sample,

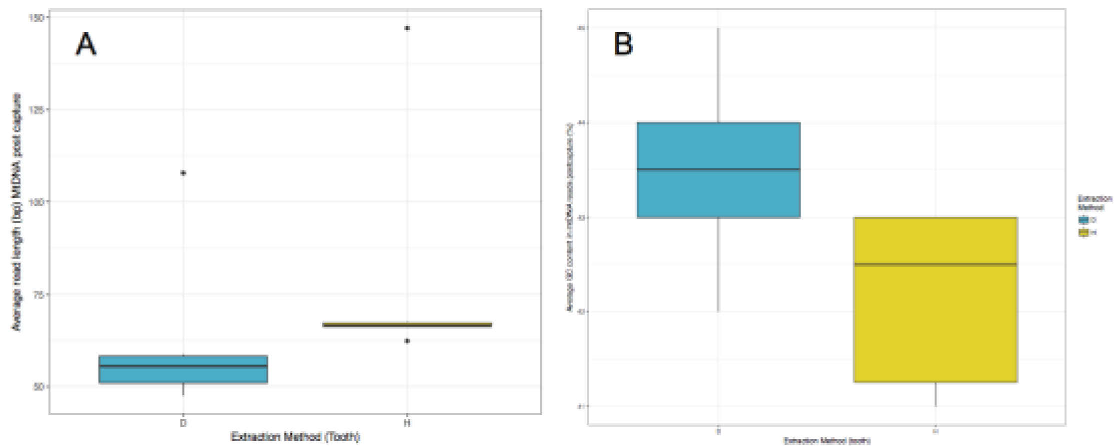


Figure 24. Distribution of average DNA fragment length (A) and average %GC content values (B) for mtDNA enriched libraries.

probabilities of C to T transitions at the first position were lower than A to G transitions at the last position (0.88 vs 0.95). This suggest less DNA damage occurred in these samples. Average $\delta_S$ values were significantly different between enriched samples extracted with different methods (Table 8). Method D samples exhibited higher $\delta_S$ values suggesting that successfully captured mtDNA fragments extracted with this method had higher probability of carrying deaminated cytosines in single strand contexts. No statistically significant differences were found in $\delta_D$ and $\lambda$ values for enriched mtDNA samples (Figure SM28). Overall, damage parameters were lower in enriched mitochondrial reads from the chimpanzee samples than the human samples regardless of extraction method. Lastly, contamination estimates performed on a subset of mtDNA enriched samples with >3x read depth and >90% coverage of the mitochondrial genome estimated most samples had over 90% proportions of authentic ancient DNA (Table SM6) (Fu et al. 2013)

## 3.5 Discussion

In this research, I explored the efficiency and performance of two extraction methods on poorly preserved ancient and historic tooth samples from tropical contexts. Experimental results suggest that both methods tested (Method D and Method H) successfully recovered degraded genetic material from tooth remains. No significant difference was observed in raw DNA yields, endogenous DNA content, library complexity (measured as distinct reads after shotgun sequencing) or postmortem damage of shotgun reads. The latter suggests that neither method is biased against recovery of degraded DNA fragments and is consistent with results previously reported by (Gamba et al. 2016) who found that ancient samples extracted with different silica based extraction methods did not have different postmortem damage patterns. Other studies that have conducted comparative analyses of shotgun reads obtained with different extraction methods have found that modifying digestion or pre-digestion wash steps also had negligible effects on DNA damage profiles (Boessenkool et al. 2016, Damgaard et al. 2015).

Significant differences between methods were observed in DNA fragment length and %GC content of shotgun sequence reads. DNA fragments recovered with Method H were, on average, 15 base pairs longer than those recovered with Method D. This difference appears to stem from increased recovery of smaller DNA fragments with the latter method and not from loss of larger DNA fragments (Figure 21). Extraction method D was designed for recovery of ultrashort DNA fragments (Dabney et al. 2013). Given that Method H conserves the binding and purification steps implemented in Method D,

short DNA fragment loss may be occurring during the pre-digestion EDTA wash or during the extended digestion and decalcification step (Figure 15).

Warinner et al. (2014) did not observe reductions in raw DNA yields after pre-extraction washes of calculus samples with EDTA. However, they did not test whether fragment length distribution was different between samples extracted with and without treatment. It is possible that implementation of the wash step has different effects in different tissues, leading to increased loss of small, endogenous DNA fragments in teeth as opposed to calculus. Previous research by (Damgaard et al. 2015, Korlevic et al. 2015, Gamba et al. 2016) did not find significant differences in average DNA fragment lengths after implementation of extended digestion steps or decontamination procedures in ancient tooth extractions. However, Boessenkool et al. (2016) observed that mean aDNA fragment lengths were smaller and GC content was higher in bone extractions performed with short versus long digestions. Although Damgaard et al. (2015) did not find differences in fragment length across a variety of digestion times, they did observe diminished recovery of endogenous aDNA with digestion steps longer than one hour. As implemented here, Method H has an extended two-part digestion step that consists of incubating bone powder for one hour in lysis buffer, followed by a second overnight digestion at 37°C. It is possible that optimization of digestion conditions such as temperature and incubation time is needed to avoid loss of small DNA fragments. Future optimization efforts may also benefit from separate library preparation and sequencing of EDTA wash and pre-digestion fractions to identify whether and at what step in the process small DNA fragments are being lost.

I observed that average GC content was at least three percentage points higher in paired tooth samples extracted with Method D versus Method H. Previous research has shown that differential DNA preservation can cause compositional bias in ancient genomes towards higher GC content (Briggs et al. 2007, Schuenemann et al. 2011, Krause et al. 2010). Higher GC content has also been correlated with lower contamination due to reduced presence of exogenous DNA (Racimo, Renaud, and Slatkin 2016). Although GC content can be strongly affected by amplification enzymes used in the library preparation process (Seguin-Orlando et al. 2015, Dabney and Meyer 2012, Aird et al. 2011), all samples in this study were amplified with the same conditions so this is unlikely to explain the observed differences in base composition between extraction treatments. At first glance these results suggest that since Method D likely allowed for higher recovery rates of GC-rich DNA, it may be better suited for ancient tropical samples. However, no significant correlation was identified between %GC and endogenous content in this investigation (Figure SM20). High GC content can also lead to low sequence coverage in aDNA due to difficulty with mapping and alignment (Krause et al. 2010, Schuenemann et al. 2011). Although this problem may be alleviated somewhat by deep sequencing and high read depths, it is possible that increased recovery of GC-rich DNA, may not necessarily lead to better results when read depth and coverage is inherently low, such as in poorly preserved tropical samples.

Lastly, there were no significant differences in the percent of distinct reads obtained from Method D and Method H samples during the shotgun sequencing run. However, extrapolation of predicted library complexity with the two best preserved samples (PI 67 and GB 7) indicated that Method H libraries would yield fewer unique

DNA fragments upon repeated sequencing experiments. Complexity analyses are highly sensitive to the amount of sequence data generated. Low amounts of reads can lead to false estimates due to uncertainty of the extrapolation (Daley and Smith 2013). This is reflected in the large size of the confidence intervals for the extrapolation curves of Method D libraries, which yielded fewer unique, mapped reads than Method H libraries (Table SM5). The large confidence intervals indicate there is uncertainty in determining the exact point at which saturation would be reached after repeated sequencing (Figure 20). But, regardless of error margins, the large difference between the complexity curves for both methods demonstrates that deeper sequencing would likely be most useful with libraries constructed from Method D versus Method H extracts.

Mitochondrial capture resulted in substantial enrichment of mtDNA sequence reads regardless of extraction method. Method D samples had significantly higher average enrichment folds, which suggests that the capture was more efficient in these libraries. However, no clear relationship was observed between pre-capture and post-capture mtDNA endogenous content. For most samples, very few mtDNA sequence reads sequenced in the shotgun library were retained in the captured library (Table SM5). Further, high mtDNA endogenous content in the shotgun libraries did not predict mtDNA endogenous content after capture, and no significant difference was detected in mean endogenous content between enriched libraries from different extraction treatments. These results suggest mtDNA capture results could be strongly influenced by sampling bias in the population of available DNA fragments when dividing library fractions for shotgun versus capture applications (Cruz-Davalos et al. 2016). There may also be an overrepresentation of mitochondrial DNA fragments relative to autosomal DNA

fragments in aDNA extracts, due to the higher number of mitochondria present in eukaryotic cells (Giles et al. 1980).

Significant between method differences in read length distribution and GC content bias identified in the shotgun libraries were retained after mtDNA capture. But, fragment length increased substantially after enrichment. As noted by (Cruz-Davalos et al. 2016, Ávila-Arcos et al. 2015), larger DNA fragments may preferentially bind to enrichment capture baits and skew read length distributions in post-capture datasets. Thus, it is expected that DNA fragments after enrichment may, on average, have larger sizes than after shotgun sequencing. In this study, the bias introduced by the capture likely amplified the differences previously observed in the shotgun data. GC content may also be skewed by capture methods. Avila-Arcos et al. (2011) found that whole genome capture in ancient maize samples was less efficient at extremely low or high GC content genomic regions. Other studies have also found a range of differences in pre and post contact GC content, including an increase (Gnirke et al. 2009), a reduction (Carpenter et al. 2013) and a narrower distribution of GC values after whole genome enrichment in ancient human samples (Ávila-Arcos et al. 2015). Therefore, significant differences in GC content between enriched samples in this experiment may stem from preservation and library amplification biases, as well as from subsequent biases introduced during the enrichment process.

Damage patterns in the mtDNA enriched samples also replicated results seen in shotgun libraries, except for mean $\delta_S$ values which were significantly higher in Method D samples. $\delta_S$ measures the probability that DNA fragments carry deaminated cytosines in single strand contexts (Jónsson et al. 2013). However, since the library preparation

and capture processes are designed to yield double stranded sequences, it is unclear whether this results in differential improvements in downstream sequence coverage, read depth or analyses for samples extracted with each method.

**3.6 Conclusion**

This study finds that Method D and Method H were similarly efficient at recovering endogenous DNA in archaeological and historic teeth from tropical contexts. But, significant differences exist in the composition of the recovered sequence reads. Method D recovers smaller aDNA fragments than Method H. Although this difference may be less important when working with well-preserved samples, the exacerbated aDNA degradation that takes place in the tropics, produces large proportions of small fragments in ancient substrates (Allentoft et al. 2012, Hofreiter et al. 2015). Therefore, my findings suggest that until further optimization of new protocols can take place, Method D continues to be the optimal choice for maximizing aDNA recovery in tropical tooth samples.

An important caveat to this finding is the small sample size of the present dataset. With a paired design of six individuals, statistical power may be too low to detect small but important differences. Future research into this topic will benefit from increased sampling. A larger dataset would further allow for finer subsetting of the data such that study results could control for differences between relatively well versus poor preserved samples and/or for differences in site-specific aDNA preservation patterns. It would be of additional interest to also include other tissue substrates, such as dental calculus or

petrous portions, into future efforts to develop methodologies tailored for tropical aDNA samples.

CHAPTER 4

# RECONSTRUCTING THE PEOPLING OF PRE-CONTACT PUERTO RICO THROUGH ANCIENT DNA ANALYSIS

## 4.1 Abstract

Native American groups exclusively occupied the Caribbean island of Puerto Rico for 4,000 years before present, until European colonization in 1493. Due to the demographic shifts that occurred after contact, the origins of these ancient populations and their contributions to the ancestry of modern Puerto Ricans are still disputed. Further, although there is archaeological evidence of continuous interaction between ancient Puerto Rican groups and other Caribbean communities, the role of genetic exchange in maintaining these social networks is currently unknown. In this chapter, I use ancient DNA to characterize migration into pre-contact Puerto Rico and to examine the extent of genetic continuity between ancient groups and modern Puerto Ricans. I sampled dental remains from 60 individuals (dated between A.D. 500–1300) collected from three archaeological sites: Tibes, Paso del Indio and Punta Candelero. Using in-solution capture and next-generation sequencing, I obtained 27 complete mitochondrial genomes (mean read depth: 18.9x) and partial autosomal genotypes from one individual. Results indicate there is a high proportion of Native American haplogroups A2 and C1 in this sample (29% and 55%, respectively). This distribution, as well as the haplotypes represented, supports a primarily South American origin for pre-contact populations, but does not exclude possible contributions from other regions. In addition, I find two identical mtDNA haplotypes in pre and post contact Puerto Rican populations that are not

shared with any other population in a dataset of 735 complete mitochondrial genomes. Lastly, I observe limited instances of mtDNA lineage sharing between coeval pre-contact island populations, suggesting that female-mediated gene flow was not essential to Pan-Caribbean social interaction. Nuclear genotypes generated from one ancient sample did not further inform these issues and repeated attempts at recovering autosomal aDNA are currently underway. Project findings contribute to a more complete reconstruction of pre-contact population history in the Caribbean and increase understanding of how ancient migration and admixture has shaped the biological and cultural diversity of modern Puerto Rican populations.

**4.2 Introduction**

This research investigates the role of migration and genetic admixture in the demographic history of pre-contact populations from Puerto Rico. The Caribbean islands, also known as the Antilles, form an archipelago that extends north to south along the Caribbean Sea (Figure 25). Puerto Rico is the smallest of the Greater Antilles, located east of Hispaniola and west of the Lesser Antilles (Rogozinski 2008). Although the Caribbean is a small geographical region, there is abundant archaeological evidence for multiple peopling events and continuous settlement of many of the islands throughout antiquity (Wilson 2007, Rouse 1992, Siegel 2005, Rodriguez Ramos 2010a, Curet 2004). Due to this, Caribbean archaeology has placed a heavy emphasis on the study of migration and its role as an agent of cultural and biological change (Siegel 2013, Rouse 1992, Curet and Stringer 2010). At the regional scale, much of this research has focused on reconstructing the original peopling of the Antilles (Curet 2004). However, the source

of the original populations, the number of migratory waves, and the extent and role of genetic admixture in this process remain disputed (Chanlatte Baik 2003, Rodriguez Ramos, Pagán Jiménez, and Hofman 2013, Keegan 2013). At the local scale, there is broad agreement that ancient island communities engaged in strong and continuous interaction (Rodriguez Ramos, Pagán Jiménez, and Hofman 2013, Hofman et al. 2011). But, identifying the precise nature and mechanisms for this interaction (e.g. trade, kin networks, spousal exchange, political alliances) remains a difficult task (Curet and Hauser 2011, Morsink 2013, Mol 2013, Laffoon 2013).

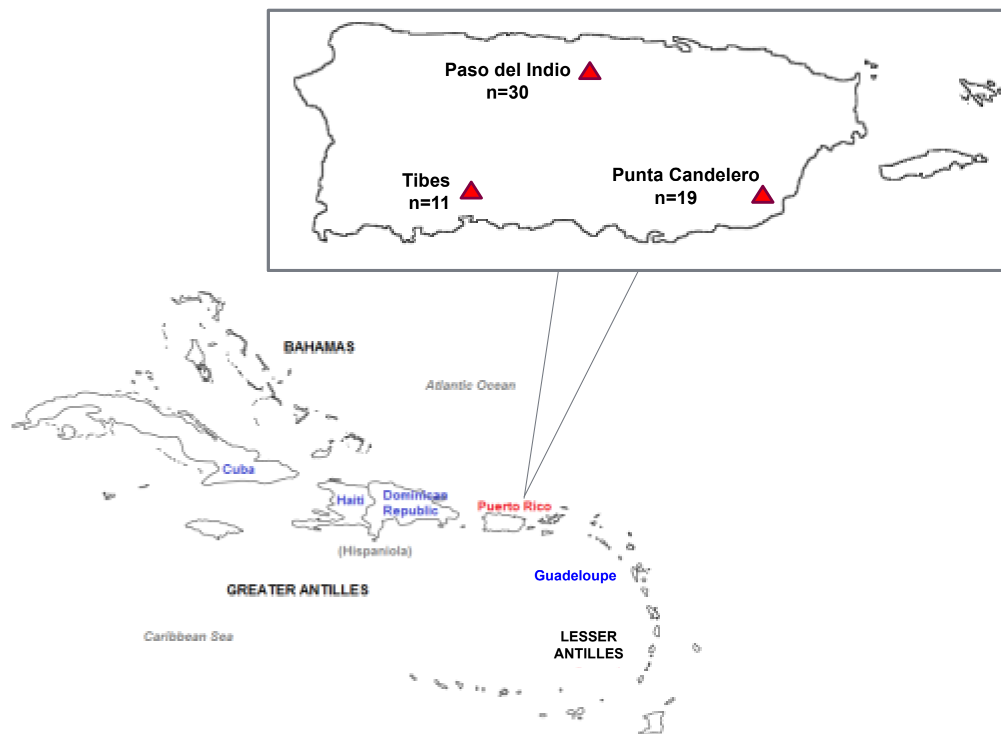


Figure 25. Map of Puerto Rico and the Caribbean islands. Islands for which pre-contact ancient mitochondrial DNA data are available are labeled in blue. Inset zooms into Puerto Rico to demonstrate approximate location of three archaeological sites from which skeletal remains were sampled. Sample size for each site is noted. Total sample size is N=60 (Images modified from WorldAtlas).

In Puerto Rico, geneticists have attempted to trace pre-contact interactions by reconstructing the genetic diversity of ancient groups through Native American genome segments found in modern, admixed Puerto Ricans (Gravel et al. 2013, Martínez-Cruzado 2002, 2010, Moreno-Estrada et al. 2013, Via et al. 2011, Vilar et al. 2014, Martínez-Cruzado et al. 2005). But, because modern populations do not retain all the genomic diversity of ancient groups, research with contemporary individuals provides limited resolution for reconstructing the scale and structure of ancient demographic events (Bolnick 2011, Pickrell and Reich 2014). This problem is exacerbated in Caribbean populations where a recent history of colonialism, slavery, and migration led to extensive genetic admixture and cultural creolization throughout the post-contact era (Knight 1997b, Bryc, Velez, et al. 2010, Moreno-Estrada et al. 2013, Hoetnik 1985). To address these limitations and build upon previous efforts, this investigation uses ancient genomics to examine directly the genetic diversity of pre-contact Puerto Rican communities and evaluate models for its population history.

### 4.2.1 *The peopling of Puerto Rico and the Caribbean: The (bio)archaeological and genetic evidence*

During the late 19th and early 20th centuries, Caribbean archaeologists focused on defining "culture areas" for human groups based on the description and classification of artifact complexes (Siegel 2013). This approach was reflected in the work of Irving Rouse, who developed a scheme of modal pottery analysis which defined the geographic and temporal boundaries of pre-contact Caribbean cultures (Rouse 1992, 1986). Rouse's model proposed that distinct cultural groups were represented by artifactual proxies and

that change in these assemblages over time resulted from migration and replacement of pre-existing populations (Rouse 1986, Cabana 2011). For instance, material evidence indicates that humans carrying chert blade technology first arrived in the Caribbean by 5000 B.C. However, by 3000 B.C., ground stone technology assemblages appear in Puerto Rico and other islands. Rouse proposed that the two technologies (labeled Casimiroid and Ortoroid), stemmed from independent migrations originating in Central and South America, respectively. As such, under this model, the arrival of the Ortoroid represented the first population transition event in the Antilles. Rouse thought these populations were nomadic hunter-gatherers but more recent work indicates that they possessed small-scale agriculture and ceramic technology (Cruxent and Rouse 1969, Rouse 1986, 1992, Reid 2014).

Rouse identified a second population transition between 400 and 500 B.C. which he labelled the Ceramic Age to differentiate it from the earlier Lithic Age. This transition was defined by the introduction of large-scale agriculture and the appearance of elaborate ceramic technology similar to that of Arawak cultures from the South American Orinoco region (Rouse 1992, 1986, Chanlatte Baik 2003). Rouse proposed that this change signaled the arrival of a new group of migrants from South America, known as the Saladoid, who displaced the first inhabitants and became the sole ancestors of all later Antillean cultures. Thus, under this model, indigenous Puerto Rican groups at the time of European contact, known as the Taino, were direct descendants of the second re-peopling event and carried little to no genetic or cultural ancestry from the initial island settlers (Rouse 1992).

Although Rouse's model continues to influence modern Caribbean archaeology, recent scholarship has challenged many of its central tenets. Questions of where the inhabitants of the islands came from, how they reached the islands, and what was the nature of their interactions continue to be a central focus of archaeological research (Keegan 2013, Siegel 2013). Many of these questions have been re-examined through novel and interdisciplinary approaches. For example, extensive radio-carbon dating surveys have found that the earliest Ortoroid and Saladoid sites from Puerto Rico pre-date the oldest sites in the Lesser Antilles (except for Trinidad), thereby challenging Rouse's model of a stepping-stone movement from South America into the Greater Antilles (Rodriguez Ramos 2010a). Additional migration routes from South America have also been proposed. Callaghan (2003) performed computer simulations of ancient maritime voyages that considered climate, maritime conditions, and pre-contact vessel technology. The results indicate that direct canoe travel from the northern South American mainland to the Greater Antilles would be possible for seafaring cultures with minimal navigation skills (Callaghan 2003, 2013). Further, findings of cultigens, lithics, and other objects from the Isthmo-Colombian region at Lithic Age sites in Puerto Rico and Virgin Islands has led researchers to consider this region as a potential source area for pre-contact settlers (Rodriguez Ramos 2013).

Alternative views to the single-origin model have also been proposed. The persistence of Lithic Age iconography and manufacturing techniques in late Ceramic Age stone tools, art, and pottery suggests that descendants of Lithic Age groups present during the transition may have admixed with the newcomers and contributed to the development of descendant populations (Keegan 2006, Keegan and Rodriguez Ramos 2005, Rodriguez

Ramos 2010a, Wilson 1999). Further, the lack of evidence for systemic conflict during the transition years has led some researchers to question the population-replacement scenario proposed by Rouse (Curet 2005). Perhaps the strongest challenge to the single-origin model stems from the discovery of distinct archaeological assemblages at two Ceramic Age sites in Puerto Rico, Punta Candelero and La Hueca-Sorcé (Chanlatte Baik 2003, Rodriguez Ramos 2010b). Multiple lines of evidence suggest that these sites were occupied by at least two different ethno-cultural groups with separate origins within South America (Chanlatte Baik 2003). This suggests that Puerto Rico, and perhaps the other Antilles, were settled by at least two separate South American populations during the Ceramic Age.

To explore these questions further, many researchers have turned to direct measurements of ancient biological variation. Overall, these studies vary in their support for the admixture versus single origin models and offer differing views of ancient inter-island mobility and interaction. For example, after A.D. 600, a period known as the Late Ceramic Age, cultural diversification took place across the Antilles (Wilson 1999, 2007). Social and political changes led to increased social stratification, the rise of chiefdoms, and the emergence of regional technological and artistic industries (Curet, Torres, and Rodríguez 2004, Rouse 1992, Chanlatte Baik 2003). Although extensive networks existed between the Antilles and the Circum-Caribbean since the Lithic Age (Hofman et al. 2011, Rodriguez Ramos 2013, 2011, Laffoon et al. 2014), material and stable isotope evidence suggest that inter-island interaction, trade and mobility expanded during the late Ceramic Age (Laffoon et al. 2013, Hofman et al. 2011, Wilson 2007). Isotopic studies with human skeletal remains from the Lesser Antilles find evidence for substantial,

mostly female-biased, inter-island adult mobility during this period (Hoogland, Hofman, and Panhuysen 2010, Laffoon and Hoogland 2012). These findings are consistent with material evidence for broad inter-island interaction and with ethnohistoric evidence for virilocal kinship and residential mobility patterns across the Antilles (Morsink 2013, Keegan and Maclachlan 1989).

In contrast, dental biodistance studies do not find strong evidence for shared morphological traits across island groups, which suggests there was no, or at least very little, inter-island admixture (Coppa et al. 2008). Instead, cranial and dental biodistance studies find that most ancient Antillean groups share morphological affinity with native South American populations. Ancient groups from Cuba, however, consistently appear as morphologically distinctive from other Antillean populations and, thus, may have a distinct origin (Coppa et al. 2008, Ross, Ubelaker, and Falsetti 2002, Ross and Ubelaker 2010, Ross 2004). Overall, these findings support a multiple migration model for the peopling of the Antilles.

Genetics research with modern Caribbean islanders provides further insight into these issues. Due to post-contact admixture, islanders have varying proportions of African, European and Native American ancestry (Benn-Torres, Kittles, and Stone 2007, Moreno-Estrada et al. 2013). These patterns differ among populations, and they are highly sex-biased and geographically structured (Marcheco-Teruel et al. 2014, Via et al. 2011, Bryc, Velez, et al. 2010). For example, modern Puerto Ricans have high proportions of European and African ancestry in autosome and Y-chromosome loci (Moreno-Estrada et al. 2013, Parra, Kittles, and Shriver 2004, Ruaño et al. 2009, Via et al. 2011, Vilar et al. 2014), but also carry large proportions of Native American mtDNA

ancestry. This Native American component has been used as reservoir for reconstructing pre-contact genetic variation (Gravel et al. 2013, Martínez-Cruzado 2002, 2010, Martínez-Cruzado et al. 2005, Moreno-Estrada et al. 2013, Vilar et al. 2014)

Native American populations have five clinally distributed mtDNA lineages, or haplogroups: A2, B2, C1, D1, and X2. Haplogroups A2 and B2 are found in high frequencies in North and Central America, C1 and D1 occur in high frequencies in South America, and X2 is only found in North America (Schurr 2004). Each of these haplogroups can be further divided into sub-haplogroups, or haplotypes. Modern Puerto Ricans carry high frequencies of mtDNA haplogroups A2 and C1. This distribution suggests a primary northern South American origin, with some smaller contributions from Central American sources (Martínez-Cruzado 2002, Martínez-Cruzado et al. 2005, Vilar et al. 2014). Based on a sample of 122 mtDNA haplotypes, Martínez-Cruzado (2010) proposed that 84% of Native American mtDNA lineages in Puerto Rico stem from ancient indigenous groups and arrived on the island through at least three major waves during the pre-contact era. The remaining 16% were of unknown origin and may have been introduced after European contact. However, studies examining autosomal loci have found that genome fragments of Native American ancestry in Puerto Ricans, Cubans, and Dominicans cluster closely with modern Amazonian groups. This suggests that the ancestors of native Caribbean groups descended primarily from South American populations and had limited genetic contributions from other sources (Gravel et al. 2013, Moreno-Estrada et al. 2013).

However, using modern genomes to reconstruct ancient population processes is problematic because modern populations may not retain all the genomic diversity of

ancient groups. Evolutionary forces such as genetic drift and natural selection affect lineage survival in descendant populations and bias reconstructions of ancient demography and population history (Bolnick 2011, Helgason et al. 2009). Recent population replacements can also mask the genetic signal of ancient populations (Bramanti et al. 2009, Haak et al. 2015). In Puerto Rico for example, historical documents indicate that native groups from other parts of the Americas were imported as slave labor during the early 16th century (Whitehead 1999, Anderson-Córdova 2005, Haslip-Viera 2006), yet their contribution to the gene pool of modern Puerto Ricans is currently unknown (Martínez-Cruzado 2010).

Ancient DNA (aDNA) research has the potential to resolve many of these issues. But, due to the constraints that tropical environmental conditions pose for aDNA preservation, most work in the region has been limited to recovery of partial mtDNA sequences. To date, ancient mtDNA has been recovered from skeletal remains excavated at Ceramic Age sites in Cuba, Hispaniola and Guadeloupe. The data indicate that populations living between 670 and 1600 AD on these islands carried high frequencies of haplogroups A2, C1, and D1. This distribution is consistent with a South American origin for these populations, but cannot conclusively rule out a Central American genetic contribution (Lalueza-Fox et al. 2003, Mendisco et al. 2015). In agreement with patterns also observed in the modern mtDNA data, all aDNA studies found high proportions of unique Native American mtDNA haplotypes, not shared across island populations. In Guadeloupe, limited lineage sharing is also found between archeological sites and is correlated with differences in local funerary practices (Mendisco et al. 2015). Taken together, this suggests that female-mediated gene flow was limited during the pre-contact

era and that restricted admixture could have contributed to inter-island cultural differentiation during the late Ceramic Age (Lalueza-Fox et al. 2001, Lalueza-Fox et al. 2003, Mendisco et al. 2015, Vilar et al. 2014, Mendizabal et al. 2008, Martínez-Cruzado 2013). However, due to the paucity of available skeletal remains and insufficient DNA preservation, sample sizes obtained for all Caribbean aDNA studies were small ($n \leq 25$), and none sequenced autosomal loci. These factors constrain the applicability of inferences gleaned from those studies to other ancient populations such as those from pre-contact Puerto Rico (Martínez-Cruzado 2010, Schurr 2010).

Thus, significant gaps remain in current understanding of the population history of ancient Puerto Rico and the Caribbean. Multiple lines of evidence disagree on the sources and number of human migrations, and there is no consensus regarding the single versus multiple migration model for the ancestry of late Ceramic Age populations. Although there is abundant archaeological evidence for extended social networks in the pre-contact Antilles, the biological evidence sampled to date is not consistent with broad inter-island gene flow. And lastly, although genetic research has been conducted with modern Puerto Ricans, the extent to which ancient populations contributed to the ancestry of extant islanders has not been fully characterized. To address these issues, this research uses ancient DNA to characterize the genetic diversity of pre-contact populations from three pre-contact archaeological sites in Puerto Rico: Punta Candelero, Paso del Indio and Tibes. Research questions for this study are:

(1) How were ancient populations in Puerto Rico related to continental Native American groups, and what may this indicate about their origins?

(2) What was the extent of gene flow in the pre-contact Caribbean, and what may this

indicate about local social interaction?

(3) How and to what extent did pre-contact indigenous groups contribute to the ancestry of modern Puerto Ricans?

## 4.3 Materials and Methods

### 4.3.1 *Sample and site information*

The site of Punta Candelero is located on a coastal peninsula in southeastern Humacao, Puerto Rico (Figure 26). It was identified and excavated between 1986-1989. The site had two successive occupation periods: 350 BC-210 A.D. and 660-1010 A.D. Each period is associated with distinct site usage patterns, and ceramic and lithic assemblages, all of which suggest a multicomponent site occupation. Household structures and general activity areas (including a central plaza built during the later period) have been identified at the site. No human burials were recovered from the early period but 106 skeletal remains were reported from late period strata, 78 of which have been identified as human (Crespo-Torres 2000, Fontanez 1991, Rodríguez López 1991).

The site of Paso del Indio is located in the alluvial plain of the Río Indio in north-central Vega Baja, Puerto Rico. It was identified and excavated between 1993-1995. Paso del Indio was non-continuously occupied from 2690 B.C. to 1440 A.D. Interruptions in habitation may have been partly due to periodic river flooding. Large quantities of post molds indicate that many household structures, and potentially one central plaza, were built at Paso del Indio throughout its occupation. One-hundred and thirty eight human skeletal remains were recovered from the site (Walker 2005). Skeletal remains from Paso del Indio and Punta Candelero are currently housed by Dr. Edwin Crespo-Torres at the

Forensic Anthropology and Bioarchaeology Laboratory at the University of Puerto Rico. For this research, thirty individuals were sampled from Paso del Indio and nineteen from Punta Candelero.

The site of Tibes is located in an alluvial terrace near the Río Portugués in southern Ponce, Puerto Rico. The site was first identified and excavated between 1975-1981. In 1982 an archaeological park, the Ceremonial Center of Tibes (CCT), was built at the site. Research and excavations resumed in 1995 and continue to the present day. Tibes was continuously occupied between A.D. 300 – 1200. In addition to habitation areas it has several monumental structures such as at least five stone lined plazas. To date, 126 human skeletal remains have been recovered from the site (Curet and Stringer 2010). Skeletal remains from Tibes are currently housed at the CCT. For this project, eleven individuals were sampled from Tibes.

In total, sixty tooth samples were obtained from human skeletal remains of the three study sites. The remains had been previously sampled by Dr. William J. Pestle (University of Miami), Dr. Edwin Crespo-Torres (University of Puerto Rico) and Dr. L. Antonio Curet (National Museum of the American Indian) as part of previous research. Forty of the 60 individuals were radiocarbon dated and assessed for collagen content and dietary isotopes by (Pestle and Colvard 2012, Pestle 2010). Sample dates span most of the Caribbean Ceramic Age: ~500 -1300 A.D. (Figure 26, Table SM1). Compliance with Native American Graves Protection and Repatriation Act regulations was not required for this research because no federally recognized tribes claim cultural affiliation to the remains, and the law has no jurisdiction in Puerto Rico (Siegel 2011, Hutt, Blanco, and Varmer 1999, Ousley, Billeck, and Hollinger 2005).

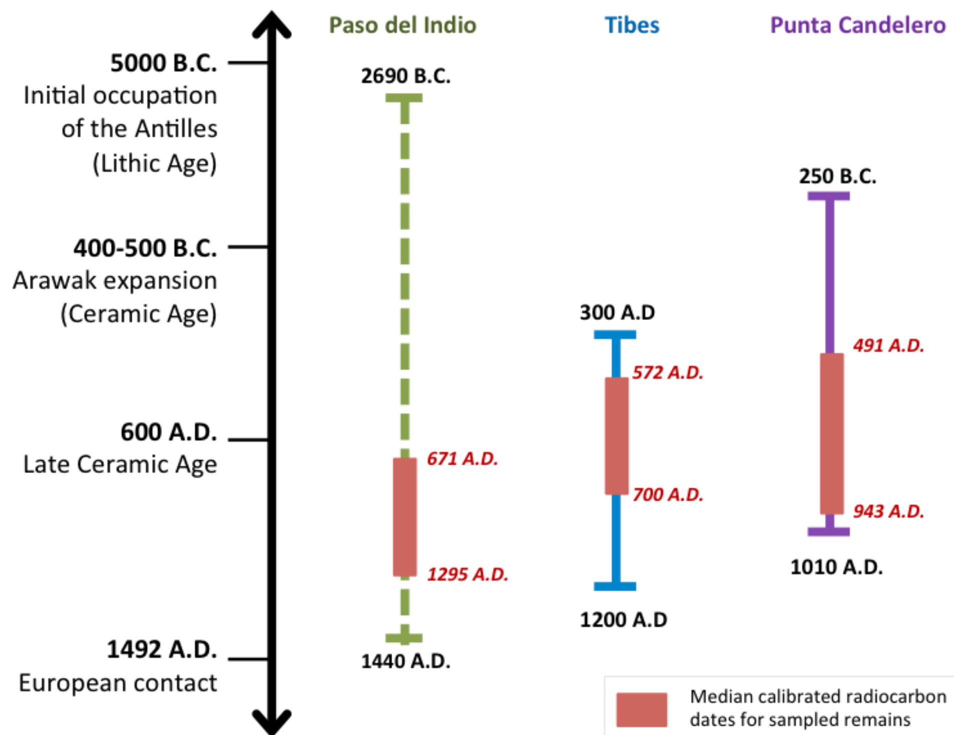


Figure 26. Simplified timeline of the pre-contact history of the Antilles demonstrating occupation periods of studied sites. Site chronologies based on published sources. Dashed lines represent the non-continuous occupation of Paso del Indio. Red boxes encompass minimum and maximum median calibrated radiocarbon dates (cal A.D.) for sampled skeletal remains.

4.3.2 *Sampling and DNA extraction*

Sampling and DNA extractions were conducted at the Arizona State University Ancient DNA Laboratory, a Class 10,000 clean-room facility. To eliminate surface contaminants and inhibitors, tooth samples were cleaned with a 1% sodium hypochlorite solution and the outer surface was mechanically removed with a Dremel tool (Rohland and Hofreiter 2007). Samples were also UV irradiated for 5 minutes on each side in a UVP CL-1000 Ultraviolet Crosslinker. Teeth were sliced transversally at the cemento-enamel junction using the Dremel. The roots were covered in aluminum foil and

pulverized by blunt force with a hammer as in (Schuenemann et al. 2011). Throughout the sampling process contamination controls such as UV irradiation of tools and work area before and between uses, full body coverings and bleach decontamination were implemented to minimize potential sample contamination (Gilbert et al. 2006, Cooper and Poinar 2000). Samples were extracted using several silica-based extraction methods for ancient DNA recovery (Table SM2), including the method developed in chapter 2 of this dissertation (Dabney et al. 2013, Rohland and Hofreiter 2007). Between 30-100 mg of dentine powder were used per extraction. DNA yields in ng/uL were measured using 1 µl of each extract through fluorometric quantification with the Qubit 2.0 High Sensitivity assay (Simbolo et al. 2013). Extraction blanks were included throughout the process. Some samples were extracted more than once to obtain more genetic material (listed in Table SM2). Fluorometric quantification values were compared to radiocarbon dates using the cor.test function in R to evaluate the relationship between sample age and overall DNA preservation.

4.3.3 *Library preparation*

Double stranded libraries were produced using 20 µl of each extract following the protocol published by (Meyer and Kircher 2010) using the Qiagen MinElute PCR purification kit instead of SPRI beads. Extraction blanks were also converted into libraries. An additional negative library control including water instead of extract was also included as a contamination control per each library batch. 1:100 dilutions of each library were quantified using Real-Time PCR (qPCR). Reactions were run in triplicate for each library in final volumes of 20 µl with the following conditions: 10 µl of 2X

Dynamo SYBR Green qPCR Master Mix with 0.3x ROX (Thermo Scientific), 1 µl of primer IS7 (5'-ACACTCTTTCCCTACACGAC-3') at 10 µM, 1 µl of primer IS8 (5'GTGACTGGAGTTCAGACGTGT-3') at 10 µM, 7 µl of $ddH_20$, and 1 µl of the library dilution. Reactions were heated to 95°C for 10 minutes, then 40 cycles of 95°C for 15 seconds and 60°C for 1 minute for 40 cycles. A final disassociation stage was added at the end of these cycles: 95°C for 15 seconds, 60°C for 15 seconds and 95°C for 15 seconds. Quantification was performed using an ABI7900HT thermocycler, and results were analyzed with SDS software. Analysis of qPCR data focused on cycle threshold values (Ct), which represent the number of qPCR cycles required for fluorescent signal to exceed background levels. Mean Ct values were averaged across all replicates per library. Non-template controls (NTC), which have no DNA, were also included in the reaction to monitor background fluorescent levels.

All libraries were double indexed and amplified following recommendations from (Kircher, Sawyer, and Meyer 2012). Primers used were Illumina specific indexing primers P5 (5'-AATGATACGGCGACCACCGAGATCTACACxxxxxx ACACTCTTTCCCTACACGACGCTCTT-3') and P7 (5'-CAAGCAGAAGACGG CATACGAGATxxxxxxGTGACTGGAGTTCAGACGTGT-3'). Unique index combinations for all samples (represented by x above) are provided in Table SM2. To increase library complexity, four 100 µl indexing reactions were performed per library. Samples processed between 2012-2014 were amplified with *Pfu*Turbo enzyme (Agilent) for 10 cycles using the following reaction conditions: 10 µl of 10X *Pfu*Turbo Buffer, 2.50 µl of 10 mM dNTPs, 1.50 µl of 10 mg/ml Bovine Serum Albumin, 2 µl of P5 indexing primer at 10000 nM, 2 µl of P7 indexing primer at 10000 nM, 1 µl of *Pfu*Turbo enzyme.

Samples processed after 2014 were indexed for 15-20 cycles with AmpliTaq Gold® enzyme (Life Technologies) following recommendations by (Seguin-Orlando et al. 2015). Reaction conditions were as follows: 9.27 µl of 10X PCR Buffer II, 3.68 µl of 10 mM dNTPs, 2.21 µl of 10 mg/ml Bovine Serum Albumin, 9.27 µl of 25 mM Gold $MgCl_2$ solution, 2 µl of P5 indexing primer at 10000 nM, 2 µl of P7 indexing primer at 10000 nM, 61.09 µl of $ddH_20$, 1.48 µl of AmpliTaq Gold® enzyme and 9 µl of DNA library. All reactions irrespective of enzyme used were heated to 95°C for 15 minutes for initial denaturation, further denaturation, annealing and elongation were performed at 95°C for 30 seconds, 58°C for 30 seconds and 72°C for 45 seconds for the determined amount of cycles. Final extension was performed at 72°C for 10 minutes, and reactions were kept at 10°C. All four aliquots of each indexed library were combined and the library was purified with Qiagen MinElute PCR purification kit following manufacturer parameters with the following modification: EB buffer was preheated to 65°C, and elution was performed in 30 µl. 1 µl of each library was used for quantification with the Qubit 2.0 Broad Range assay. Purified libraries were further diluted to a factor of 1:1000 and quantified with the KAPA Library Quantification kit (Kapa Biosystems) following manufacturer instructions on the ABI7900HT thermocycler. Fragment analysis of the indexed libraries was performed using the DNA 1000 assay on the Agilent 2100 Bioanalyzer.

Libraries initially indexed for 10 cycles which had low post purification concentrations were re-amplified to obtain sufficient DNA for mitochondrial capture (300-500 ng). Libraries were divided into four 100 µL aliquots as above. Re-amplification conditions were: 10 µl of 10X Accuprime *Pfx* reaction mix, 3 µl of IS5

primer at 10 µM, 3 µl of IS6 primer at 10 µM, 76 µl of ddH2O, 1 µl of Accuprime™ *Pfx* enzyme and 7 µl of DNA library. Reactions were heated to 95°C for 2 minutes for initial denaturation; further denaturation, annealing and elongation were performed at 95°C for 15 seconds, 60°C for 30 seconds and 68°C for 1 minute for 7-13 cycles. Final extension was performed at 68°C for 5 minutes, and reactions were then kept at 4°C. Subsequent purification and quantification was performed as detailed above.

4.3.4 *Mitochondrial target enrichment capture and Illumina sequencing*

Shotgun libraries were pooled in equimolar amounts up to 2 µg per pool. In-solution and targeted enrichment for the complete mitochondrial genome was performed as in (Maricic, Whitten, and Paabo 2010) with modifications as in (Ozga et al. 2016). Enriched libraries (including blanks) were sequenced along with other samples on multiple runs of the Illumina MiSeq using v2 2 x150 bp chemistry at the DNASU Sequencing Core at Arizona State University. Some enriched libraries were captured and sequenced more than once to increase coverage.

4.3.5 *Whole genome target enrichment capture and Illumina sequencing*

Whole genome enrichment was performed at Stanford University in Dr. Carlos D. Bustamante's genetics laboratory. Thirteen indexed and amplified libraries were screened by shotgun sequencing along with other samples on the Illumina NextSeq 500 (2 x 75 bp reads). The eight samples with the highest endogenous content (see below) were chosen for whole genome in-solution capture (WISC) which was performed as in (Carpenter et al. 2013) with the following modifications. RNA bait libraries were created from a pool

of human genomic DNA from three male individuals from the Coriell Hapmap populations (MKK, JPT and CEU). These individuals were chosen to maximize sequence diversity and obtain Y-chromosome baits. Axygen beads (1.8x volume) were used for bait library purification and xGEN ® LNA Adaptor blockers for Illumina TruSeq were used to prevent non-specific binding to the baits. WISC enriched libraries were sequenced on all four lanes of the NextSeq using same chemistry as listed above.

4.3.6 *Read mapping and processing*

Illumina sequence reads were merged and adapters trimmed using SeqPrep (https://github.com/jstjohn/SeqPrep) with a minimum overlap of 11 base pairs (bp) and a minimum length threshold of 30 bp. Read quality was assessed pre and post-merging with FastQC v.0.11.3 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For mitochondrial enriched libraries, reads were mapped to the revised Cambridge Reference Sequence (rCRS NCBI Reference Sequence: NC_012920.1). Reads recovered from shotgun and WISC enriched libraries were mapped to the hg18 (NCBI Build 36.1) reference with the mitochondria replaced by the rCRS (Andrews et al. 1999). For shotgun and enriched libraries sequenced on the NextSeq, data from multiple lanes were merged before adapter removal and mapping using the cat command in Unix. Post-capture mtDNA endogenous content values calculated from the reads were compared to radiocarbon dates using the cor.test function in R to evaluate the relationship between sample age and endogenous DNA recovery (post-enrichment).

Mapping was performed using BWA v. 0.7.5 (Li and Durbin 2009) with seeding disabled (-1 1000) and edit distance increased (-n 0.01) to improve mapping accuracy of

ancient DNA reads (Schubert et al. 2014). Sequence reads were filtered with SAMtools v. 0.1.19 (Li et al. 2009). Reads with mapping quality ≥ Q30 were retained for mitochondrial enriched libraries and ≥ Q25 for whole genome enriched samples. Duplicates were removed with the rmdup option in SAMtools, and reads mapping to more than one location were discarded by controlling for XA, XT and X0 tags. Damage patterns were characterized and read quality scores were rescaled with mapDamage v.2.0.2 (Jónsson et al. 2013, Ginolhac et al. 2011). Summary statistics such as mean read depth, standard deviation of read depth and percent of reference sequence covered were estimated on the rescaled BAM files using Qualimap v.2.2.1 (Okonechnikov, Conesa, and García-Alcalde 2016). BAM files were visualized with Tablet v.1.16.09.06 (Milne et al. 2013) and Geneious v.7.0.6 (Biomatters).

Sequence reads from enriched mtDNA libraries sequenced over multiple runs were merged after duplicate removal using SAMtools (Table SM4). Before merging, read group information was added to each sample replicate using the *AddOrReplace ReadGroups* module from Picard v.2.01 (http://broadinstitute.github.io/picard). After merging duplicate removal was repeated and damage patterns as well as summary statistics were recalculated as above.

For all mtDNA enriched libraries, mitochondrial SNP variants were called using SAMtools *mpileup* on the rescaled BAM files. Ploidy was set to 1 using the sample option. Variant calls were output in VCF format using the bcftools package. MtDNA haplogroup assignment was performed in HaploGrep 2.0 (Weissensteiner et al. 2016). Confirmation of haplogroup defining mutations and Haplogrep assignments was

performed manually with reference to Phylotree mtDNA tree Build 17 (van Oven and Kayser 2009).

4.3.7 *Contamination controls (mtDNA)*

Bayesian contamination estimates per sample were performed using the mitochondrial enriched reads with contamMix (Fu et al. 2013). This analysis provides an estimate of the proportion of contaminant DNA present in the sample reads and the proportion of authentic aDNA. A set of 311 human complete mtDNA genomes were used as potential contaminant sources. Insufficient read depth or coverage of the reference sequence can result in inaccurate contamination estimates with this method. Therefore, only samples with >3x read depth and >90% coverage of the mitochondrial genome were included in contamination assessments (Table SM3). BAM files for these samples were filtered using the program PMDtools (Skoglund et al. 2014) to retain only reads showing signs of aDNA damage (threshold > 3). Summary statistics were recalculated, and damage pattern assessment and haplogroup assignment were repeated for these samples after filtering.

4.3.8 *Consensus files and sample selection*

Twenty-seven enriched mtDNA samples with >3x read depth and >93% coverage of the mitochondrial genome were selected for population genetics analyses (Appendix I). Reads for seven of these samples were combined across multiple sequencing runs. For one sample, endogenous reads were recovered after contamination filtering with PMDtools. MtDNA consensus sequences for the 27 samples were produced from the

coverage filtered VCF files using the *vcf-consensus* module within VCFtools v.0.1.15 (Danecek et al. 2011). This allows for export of variant sites only and maintains all other sites identical to the reference. Consensus files were revised manually in Geneious and exported in fasta format. Variant sites with no coverage were designated as N. For WISC enriched libraries, reads mapping to the mitochondrion were subset from the whole genome alignments and variant calling and haplotyping proceeded as described above. This allowed for independent confirmation of lineage assignment for samples captured with both methods. Deamination damage pattern plots generated for the mtDNA reads of the 27 selected samples are available in Figure SM2.

#### 4.3.9 *MtDNA comparative populations*

Complete mitochondrial sequences were collected from the literature for 735 ancient and modern Native Americans or admixed individuals of Native American descent from across the Americas. In addition, sequences from 81 admixed Puerto Ricans of Native American mitochondrial ancestry were also included for a total dataset of 816 complete mtDNA genomes. Individual sequences were classified into 15 groups for inter-population analyses (Table SM7). Control region sequence data were also obtained from the literature for individuals with Native American mtDNA ancestry from Cuba, Dominican Republic and Puerto Rico, and members of the First People's Community (FPC), Trinidad and the Garifuna community in St. Vincent (see Table SM7 for references). Lastly, ancient mtDNA sequences were collected from individuals excavated from pre-contact sites in Cuba, Dominican Republic and Guadeloupe. Except for Puerto Ricans, complete mtDNA sequences were not available for modern or ancient Caribbean

populations so analyses including these data were restricted to the mtDNA control region (see below). Multiple alignments were prepared to account for differences in the extent of sequence data available from each dataset. For comparative analyses including just modern Caribbean populations, sequences were trimmed to a common region between positions 16024-16391. For comparative analyses including just ancient Caribbean populations, sequences were subset to positions 16056-16400. Joint median network analyses, including ancient and modern Caribbean populations, were conducted with variant sites located between positions 16056 – 16391.

For all samples, phylogenetically uninformative sites such as indels or poly-C stretches at positions 309, 315, 515-522, 16182-16183 or 16193 and mutational hotspot 16519 were excluded from analysis as recommended by (van Oven and Kayser 2009). The deletion at position 3107 was also excluded since this is a placeholder for maintaining the order of the reference mtDNA sequence. All sequence alignments were performed using the command line version of MAFFT v7.244 (Katoh and Standley 2013) and trimming was performed in Geneious v.7.0.6 (Biomatters). Data format conversions were performed using PGD Spider (Lischer and Excoffier 2012). Models of sequence evolution for each dataset were calculated using the Model Selection Tool in MEGA6 (Tamura et al. 2013).

#### 4.3.10 *MtDNA sequence and statistical analyses*

Statistical analyses of mtDNA data were performed to determine: (1) the relationship between pre-contact Puerto Rican populations (PC-PR) and continental indigenous groups, (2) the extent of gene flow and genetic differentiation across the pre-

contact Caribbean, and (3) the relationship between PC-PR communities and modern Puerto Ricans. To characterize patterns of genetic variation, intra-population diversity measures were calculated for PC-PR and for each comparative population using DnaSP v.5.10.1 (Librado and Rozas 2009). These included the number of haplotypes in each group (*h*), and nucleotide ($\pi$) and haplotype diversity (*Hd*) (Nei 1987, Nei and Miller 1990). The two latter measures estimate the probability that two randomly chosen haplotypes or nucleotides are different within a sample, respectively. For complete mtDNA genome sequences, measures were calculated across the complete molecule. To assess the effects of large differences in sample size, all populations with complete mtDNA data were randomly downsampled to 20 individuals and diversity analyses were re-calculated, as in (Martínez-Cruzado et al. 2005).

To examine patterns of inter-population differentiation and gene flow, *Fst* measures of differentiation across sub-populations were calculated for pairwise population matrices in Arlequin v.3.5 using the Tamura-Nei distance method with a gamma correction of 0.26 (Wright 1951, Excoffier and Lischer 2010, Wright 1978, Meyer, Weiss, and Von Haeseler 1999). Significance values for each pairwise comparison were calculated after 500 permutations. A significant result indicates that the null hypothesis of no difference between populations is rejected. Additionally, exact tests of population differentiation were conducted to test the null hypothesis that haplotypes are randomly distributed across all populations (panmixia). The test was run on the Genepop web server (http://genepop.curtin.edu.au/index.html) (Rousset 2008) using a Markov chain with 10,000 steps. Results from both tests were evaluated at a significance threshold of $P<0.05$. The complete mtDNA pairwise *Fst* distance matrix comparing PC-

PR to continental Native American groups was used as input for metric multidimensional scaling (MDS). MDS was performed in R using the *smacofSym* function from the smacof package (de Leeuw and Mair 2009). In addition, *Fst* heat maps were produced using the ggplot2 package to visually represent genetic distance across populations included in all comparisons. Population groups with less than 10 individuals were eliminated from this analysis. The Waorani population was also excluded from this analysis because samples published consisted of only one haplotype and therefore do not represent population-wide diversity (Table SM7).

To detect diachronic changes in genetic diversity in PC-PR, complete mtDNA sequences were divided into subsets by coeval radiocarbon dates (undated samples were excluded from this analysis). The ideal number of clusters (K) was estimated from the distribution of radiocarbon dates using the *pamk* function from the *fpc* R package (Hennig 2015). Using this value of K, data subsets were identified by hierarchical cluster analysis of radiocarbon date distributions using a custom R script (Ward 1963) (Figure SM1). A multicomponent Fisher's test with Bonferroni correction was used to test for a significant difference in haplogroup distribution across temporal groups. This was performed using the *fisher.multcomp* function from the RVAideMemoire package in R (Hervé 2016). Power for each comparison was estimated after 1000 simulations using the *power.fishertest* function from the statmod R package (Giner and Smyth 2016). Significance thresholds for all tests were set at $P<0.05$.

*Fst* and exact genetic differentiation tests were repeated comparing each temporal subset to all reference populations, including the complete PC-PR dataset, in the same manner as listed above. The relationship between temporal and genetic distance was

further assessed using Mantel tests as in (Mata-Míguez et al. 2012, Jansson et al. 2014). Pairwise *Fst* values generated with Arlequin were used to represent genetic distance and temporal distances were calculated as the Euclidean distance between average radiocarbon dates per cluster (median calAD). A Mantel test comparing the two matrices was run for 10,000 permutations with a significance threshold of $P<0.05$ using the *mantel.test* function in the R *ape* package (Paradis E. and Strimmer 2004).

To examine mtDNA variation and lineage sharing across the pre and post-contact Caribbean, and between PC-PR and continental indigenous groups, median networks were created with Network 5.1 (Fluxus). Median networks represent all possible evolutionary relationships between sequences and contain all most parsimonious phylogenetic trees. (Bandelt et al. 1995, Bandelt, Forster, and Rohl 1999, Mardulyn 2012). Haplogroup networks were conducted comparing PC-PR lineages A2, C1 and D1 to comparative populations. Mutations were identified with respect to the rCRS. Exploratory networks resulted in high levels of reticulation, therefore a weighing scheme was applied for the final round of analysis. For HVR-I networks, sites with high mutation rates as determined by (Meyer, Weiss, and Von Haeseler 1999) were weighted as in (Mata-Míguez et al. 2012). For complete mtDNA networks transversions were weighted three times as high as transitions as recommended by the Network manual. Several values of epsilon parameter were also explored but none resulted in different results than those obtained with the default of epsilon = 10. Coalescence times for several lineages were estimated from complete mtDNA networks with Network using the complete mtDNA substitution rate of one mutation every 3,624 years (Soares et al. 2009).

4.3.11 *Autosomal reads additional processing*

For population genetics analyses, autosomal genotypes obtained were compared to 938 individuals from the Human Genome Diversity Panel (HGDP) (Rosenberg 2006, Rosenberg et al. 2002). Before merging with aDNA sequence data, the reference dataset was converted to Plink format using custom UNIX scripts. Following methods outlined in (Skoglund et al. 2012), one allele at each site was randomly sampled for every individual in the panel, making each site homozygous for the drawn allele. This process effectively haplodizes the reference panel and mimics the random sampling process of low coverage aDNA reads in the sample.

Only one of the eight WISC enriched samples had high endogenous content post-capture: PI 420 (Table SM6). Genome coverage for this sample was estimated using the *genomecov* option in BEDTools v2.26 (Quinlan and Hall 2010). The sample BAM file was subset to include just those variant positions present in the HGDP reference panel. Although this approach may reduce representation of true variants, it also reduces the possibility that sequencing or damage errors are included in the analysis. Subsetting was done using SAMtools mpileup with –l option specifying a list with all variant sites present in the HGDP panel. Ancient reads were then filtered again such that only bases with BAQ > 20 and reads with mapping quality > 30 were retained. For sites covered by only one read this approach ensures only high quality bases are retained. A custom Python script was used to draw a random read for those sites with read depth > 1. This process ensured that haploid genotypes were obtained for all autosomal positions in the sample BAM. As in (Schroeder et al. 2015), a second dataset was produced following the same process but excluding all positions with T to C or A to G transitions in the ancient

sample with relation to the reference. This filter results in exclusion of all sites that could represent deamination damage. All population genetics analyses were conducted on both the raw and damage filtered datasets. The ancient autosomal genotypes were merged with the reference panel using Plink v.1.07 (Purcell et al. 2007). 11,742 sites were retained after merging for the raw dataset and 10,754 sites were retained for the damage filtered dataset. Lastly, the merged dataset was filtered for modern contamination using PMDtools (Skoglund et al. 2014) by selecting only reads with post-mortem damage scores over 0 (5,492 SNPs retained) and 3 (3,249 SNPs retained) respectively. In total, four datasets of merged genotypes were produced.

4.3.12 *Autosomal genotypes statistical analyses*

Principal components analysis was performed on the raw, damage and contamination filtered merged datasets with smartpca from the EIGENSOFT 6.0.1 package (Patterson, Price, and Reich 2006). Unsupervised global ancestry clustering analyses were conducted on the first merged datasets using ADMIXTURE v1.22 (Alexander and Lange 2011, Alexander, Novembre, and Lange 2009). This program estimates global ancestry proportions by testing data fit to an admixture model where K user defined populations have contributed to an admixed individual's genome. The analysis is 'unsupervised' because population labels are not assigned a priori. Instead groupings are determined from the data themselves (Liu et al. 2013). This analysis was conducted after thinning the merged dataset for linkage disequilibrium (LD) using the indep-pairwise option in Plink (--indep-pairwise 50 10 0.1). This command marks for removal SNPs with a pairwise $r^2 > 0.1$ within a 50 SNP sliding window, shifted forward

by 10 SNPs at a time. After the pruning step, 11,090 SNPs remained. Plotting of PCA and ADMIXTURE results was done in R.

4.3.13 *Computational resources and R packages*

This research was completed using resources from the ASU High Performance Computing Saguaro environment and the Stanford University SCG Genomics Clusters. All custom scripts written for this project will be available on GitHub after project completion. All plots and figures were generated in R 3.2.4 using the ggplot2 (Wickham 2009), scales (Wickham 2016a) and reshape 2 (Wickham 2007) packages or with R base graphics. Correlations were calculated with R base function cor.test. Other R packages used throughout this research include Matrix (Bates and Maechler 2016) and RColorBrewer (Neuwirth 2014). Python scripts were written in Python v.2.7.11.

## 4.4 Results

4.4.1 *MtDNA enriched samples: aDNA preservation, enrichment success and sample authenticity*

All samples collected for this research (N=60) were subjected to target enrichment capture with the aim of recovering moderate to high quality complete mitochondrial genomes. Nine samples were identified as having potential contamination based on low proportions of estimated authentic reads, unusually large average fragment lengths, unusual damage patterns or non-Native American haplogroups. Reads from these samples were subjected to additional contamination filters and reanalyzed. Twenty-seven

of the 60 individuals yielded mtDNA data with sufficient read depth (>3X) and genome coverage (>93%) for confident variant calling and haplotype assignment (Appendix I) . This represents 45% of the total sample. Twenty-one of these individuals had available radio-carbon dates. Mapping statistics for all mtDNA enriched samples are available in Table SM3. The final Haplogrep HSD file output listing all variant sites per individual haplotype is in Table SM16.

Successful aDNA recovery rates varied per site (Figure 27). Samples from Punta Candelero and Paso del Indio had similar rates of ancient mtDNA recovery, but 80% of samples from Tibes did not generate sufficient sequence reads to pass the depth and coverage thresholds specified above. This difference may stem from the older average age of the remains from Tibes relative to the other two sites. No significant correlation was found between sample age (measured as median calAD) and raw DNA yields ($r=-0.046$, $t=-0.31$, $df=46$, $p=0.755$). However, a slight positive correlation was identified when comparing radiocarbon dates and post-enrichment mtDNA endogenous content ($r=0.36$, $t=2.63$, $df=46$, $p=0.015$) (Figure SM3). These results indicate that younger samples yielded higher endogenous content after mitochondrial capture than older remains.

The authenticity of the human reads obtained after mitochondrial enrichment was assessed by examining damage patterns typical of ancient DNA (Briggs et al. 2007, Dabney, Meyer, and Pääbo 2013, Ginolhac et al. 2011, Jónsson et al. 2013). Because of post-mortem molecule breakdown and cytosine deamination, authentic aDNA exhibits short lengths (<200 bp) and high rates of C to T and A to G transitions at the 5’ and 3’ ends of DNA fragments. These patterns were evident in the post-capture reads

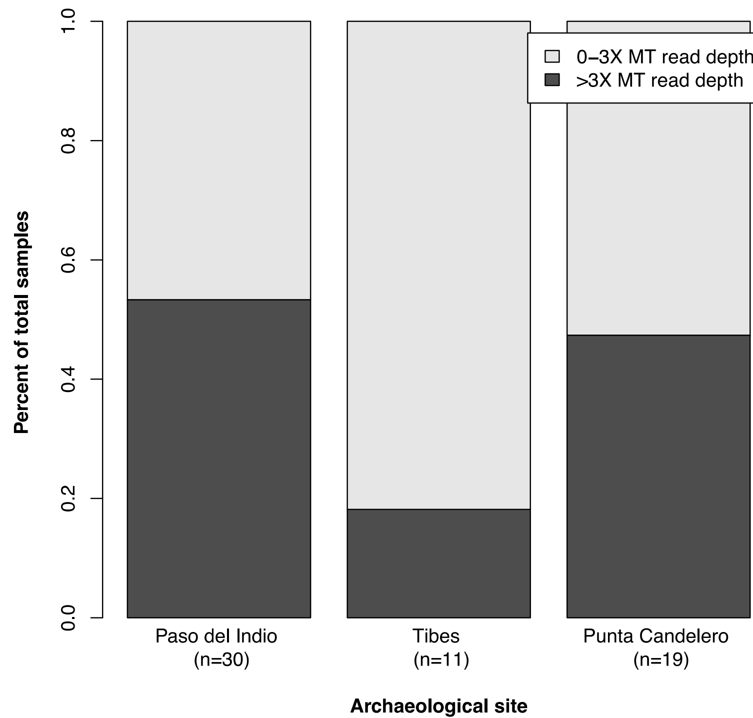


Figure 27. Per site comparison of recovery rates for complete mtDNA data after enrichment. Samples were included in subsequent analyses when read depth >3X and genome coverage >93%. MT = mtDNA.

obtained for all 27 of the selected samples (Appendix I and Figure SM2). Reads obtained from sample PC443 were contamination filtered due to initial assignment to a non-Native American haplogroup (Tables SM3 and SM5). Only sequence reads with high post-mortem damage rates (PMD score >3) were retained to ensure removal of potential contaminant sequences introduced by modern and putatively undamaged DNA sources (Skoglund et al. 2014). The proportion of endogenous mtDNA reads retained after data processing was estimated using contamMix and found to be >70% for each of the samples listed in Appendix I (Fu et al. 2013). Finally, average fragment length for all samples, irrespective of site, was approximately 62 bp. This size is similar to that

obtained in previous aDNA research with tropical Caribbean remains (Schroeder et al. 2015).

#### 4.4.2 *MtDNA diversity in pre-contact Puerto Rico*

Three of the five characteristic Native American mtDNA haplogroups were observed in the pre-contact Puerto Rico (PC-PR) sample: A2, C1 and D1 (Appendix I). Haplogroups B2 and X2a were not identified among the remains. When considering complete mtDNA variants, PC-PR individuals carried 17 unique mtDNA haplotypes, 85% of which were classified within haplogroups A2 and C1. Only two lineages were shared across multiple individuals. Haplogroup C1b2 accounted for 44% of all mtDNA lineages in PC-PR. Of the 12 individuals with this lineage, five had poor read coverage at one to six diagnostic positions which were set as Ns during variant calling (Table SM16). However, since all other diagnostic variants were present, it is reasonable to assume that these individuals fall within the C1b2 clade. When considering just HVR-I sites, PC-PR lineage variation is collapsed into nine haplotypes.

Intra-population diversity statistics indicate that the PC-PR population has lower levels of haplotype and nucleotide diversity than most comparative Native American groups, except for the Surui, Karitiana and Pima (Table 9). The low diversity of the PC-PR sample remained constant after re-calculating measures with all populations randomly down sampled to 20 individuals (Table SM8). It must be noted that diversity statistics for some continental groups may be inflated in this analysis due to grouping of the comparative data (Table SM7). Two additional comparisons restricted to the HVR-1 region and including ancient and modern Caribbean populations also found that PC-PR

had low haplotype and nucleotide diversity. Only the sample from the Trinidad FPC had lower diversity values (Table 10). For admixed Caribbean populations, this analysis was restricted to Native American mtDNA lineages.

Table 9. Diversity summary statistics calculated from complete mtDNA data.

| Population | n | S | Eta | Hap | Hd | VarHd | Pi |
|---|---|---|---|---|---|---|---|
| Karitiana | 24 | 25 | 25 | 4 | 0.5980 | 0.0033 | 0.0005 |
| Pima | 26 | 79 | 79 | 7 | 0.5200 | 0.0131 | 0.0007 |
| Surui | 23 | 41 | 41 | 3 | 0.3120 | 0.0131 | 0.0007 |
| **PC-PR** | **27** | **84** | **84** | **17** | **0.8690** | **0.0040** | **0.0012** |
| Chippewa | 13 | 67 | 68 | 12 | 0.9870 | 0.0013 | 0.0013 |
| Maya | 25 | 136 | 136 | 18 | 0.9700 | 0.0004 | 0.0016 |
| Zapotec | 88 | 239 | 240 | 57 | 0.9820 | 0.0000 | 0.0017 |
| Mazahua | 25 | 133 | 133 | 19 | 0.9670 | 0.0006 | 0.0019 |
| PC-Andes | 79 | 272 | 273 | 69 | 0.9960 | 0.0000 | 0.0019 |
| SouthCone | 106 | 291 | 292 | 97 | 0.9980 | 0.0000 | 0.0019 |
| PuertoRico | 81 | 169 | 169 | 39 | 0.9510 | 0.0002 | 0.0020 |
| MexAmerican | 215 | 654 | 659 | 213 | 1.0000 | 0.0000 | 0.0021 |
| Amazonia | 34 | 210 | 211 | 34 | 1.0000 | 0.0001 | 0.0022 |
| Andean | 9 | 113 | 113 | 9 | 1.0000 | 0.0027 | 0.0023 |
| N SouthAmerica | 22 | 179 | 179 | 21 | 0.9960 | 0.0002 | 0.0023 |

S = Number of segregating sites, Hap = Number of haplotypes, Hd = Haplotype diversity, Pi = Nucleotide diversity, average number of nucleotide differences per site between sequences.

4.4.3 *Inter-population differentiation and gene flow*

When considering complete mtDNA haplotypes, the exact test of population differentiation did not find significant differences between individuals from Paso del Indio, Punta Candelero or Tibes ($p>0.05$). Although sample sizes per individual site are low (especially for Tibes), these results suggest there is no inter-site genetic structure. Subsequent analyses were conducted assuming mtDNA lineages from intra-island communities were not strongly differentiated from each other and could be considered as one population. Exact tests identified significant differences between PC-PR and five of

the 15 comparative populations: Chippewa, Pima, Mexican Americans, Surui, Karitiana and Andeans (including Quechua and Cayapo speakers) (Table SM9). The null hypothesis could not be rejected for the remaining comparisons.

Table 10. Diversity summary statistics calculated from HVR-1 data (16024-16391).

| Population | n | S | Eta | Hap | Hd | VarHd | Pi |
|---|---|---|---|---|---|---|---|
| *PC-PR versus modern Caribbean* | | | | | | | |
| **PC-PR** | **27** | **12** | **12** | **9** | **0.698** | **0.00679** | **0.00631** |
| St Vincent | 14 | 13 | 13 | 7 | 0.868 | 0.00353 | 0.01075 |
| Trinidad | 11 | 8 | 8 | 2 | 0.545 | 0.00522 | 0.01186 |
| Cuba | 81 | 58 | 58 | 38 | 0.966 | 0.00005 | 0.01587 |
| DomRep | 82 | 40 | 41 | 29 | 0.924 | 0.00036 | 0.0173 |
| Puerto Rico | 25 | 36 | 36 | 24 | 0.997 | 0.00016 | 0.01886 |
| *PC-PR versus ancient Caribbean* | | | | | | | |
| **PC-PR** | **27** | **11** | **11** | **9** | **0.698** | **0.00679** | **0.00652** |
| PC-DomRep | 19 | 14 | 14 | 11 | 0.918 | 0.00167 | 0.00831 |
| PC-Cuba | 15 | 12 | 12 | 10 | 0.943 | 0.00162 | 0.00966 |
| PC-Guadeloupe | 10 | 13 | 13 | 9 | 0.978 | 0.00292 | 0.01346 |

S = Number of segregating sites, Hap = Number of haplotypes, Hd = Haplotype diversity, Pi = Nucleotide diversity, average number of nucleotide differences per site between sequences.

In comparisons including PC-PR, the three lowest *Fst* sub-differentiation values were observed between this group and indigenous populations from northern South America, Amazonia, and extant Puerto Ricans of Native American mtDNA ancestry (Table SM10). Some similarity is also apparent between PC-PR and Mexican Americans. However, the null hypothesis of no difference between populations was rejected for all comparisons in the permutation test. *Fst* distances between populations are visually represented with MDS in Figure 28A. The Surui and Karitiana were highly differentiated from all populations in this dataset and therefore were eliminated from the plot to allow

for better visualization of clustering patterns. PC-PR falls far from the central cluster of South American populations in dimensional space. Additionally, PC-PR is also separated from modern Puerto Ricans. This suggests that despite having relatively low *Fst* differentiation values, mtDNA haplotype composition differs between the pre and post contact population. This is especially noticeable when the distance between the two data points is compared with the closeness exhibited by ancient and modern Andean populations on the MDS plot and their extremely low subdifferentiation values in the pairwise *Fst* matrix. Lastly, the MDS plot recapitulates the stark differentiation between the mtDNA pools of PC-PR, and North American groups such as the Chippewa and Pima. These observations are limited by a relatively poor data fit for the MDS plot, as shown by the stress value >0.05 (Kruskal and Wish 1978). Therefore, *Fst* values were also visualized as a heatmap matrix in Figure 28B, where darker colors represent increased genetic differentiation (higher *Fst* distance).

Two additional pairwise *Fst* matrices were calculated including HVR-I haplotypes from PC-PR and several Caribbean groups. The first analysis, which included just ancient populations, found very small *Fst* differentiation values between ancient communities (Table SM11). However, a significant difference was observed between PC-Dominican Republic (PC-DomRep) and several other populations, including PC-PR. The PC-DomRep mtDNA pool has high rates of C and D haplotypes but no A lineages (Lalueza-Fox et al. 2001). This distribution contrasts with that observed in other ancient Caribbean populations which have varying frequencies of A2, C1 and D1 haplogroups (Lalueza-Fox et al. 2003, Mendisco et al. 2015).

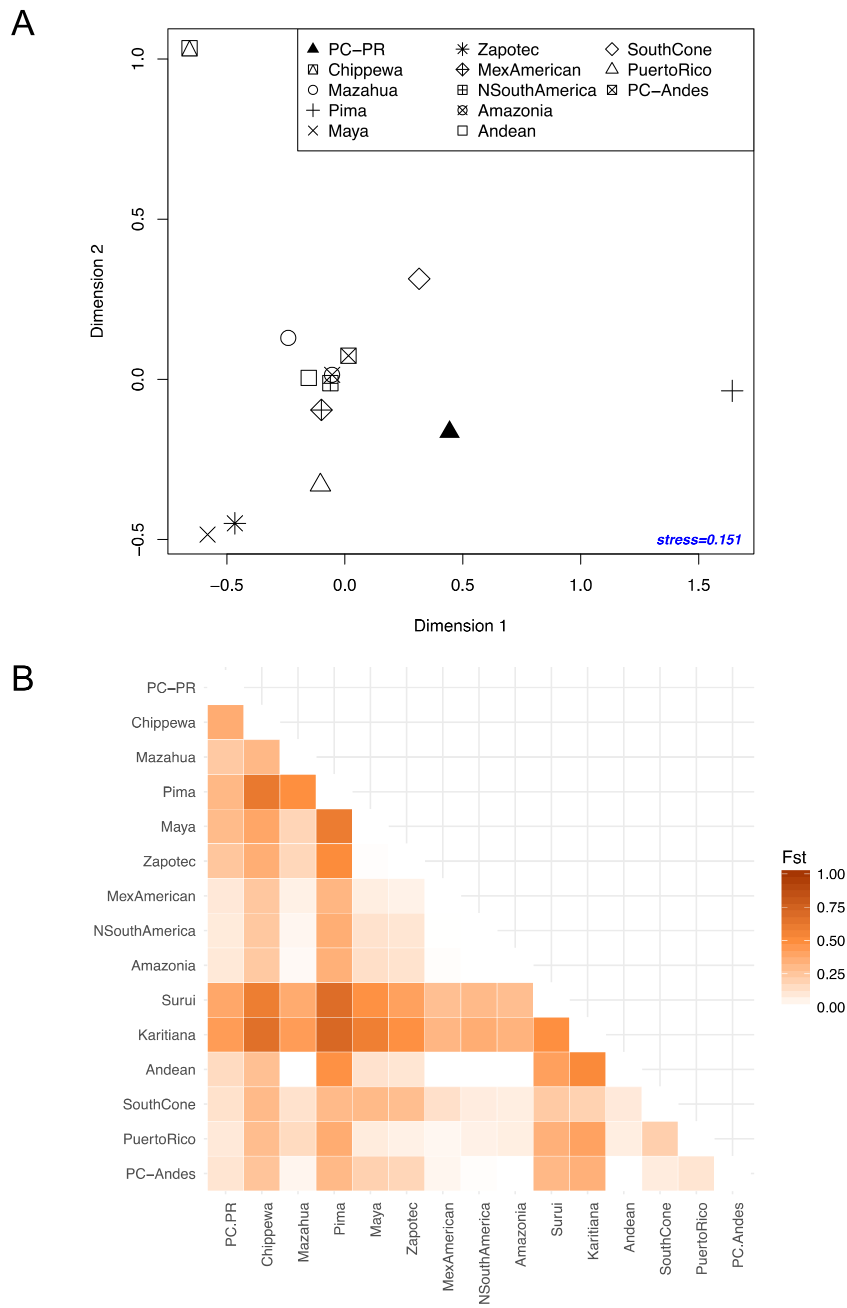


Figure 28. Fst genetic distances from complete mtDNA sequence data. (A) MDS plot in two dimensions. Surui and Karitiana were excluded from the plot due to falling extremely far from other populations. (B) Heatmap matrix of raw pairwise *Fst* values. Darker colors represent increased genetic differentiation. PC = pre-contact.

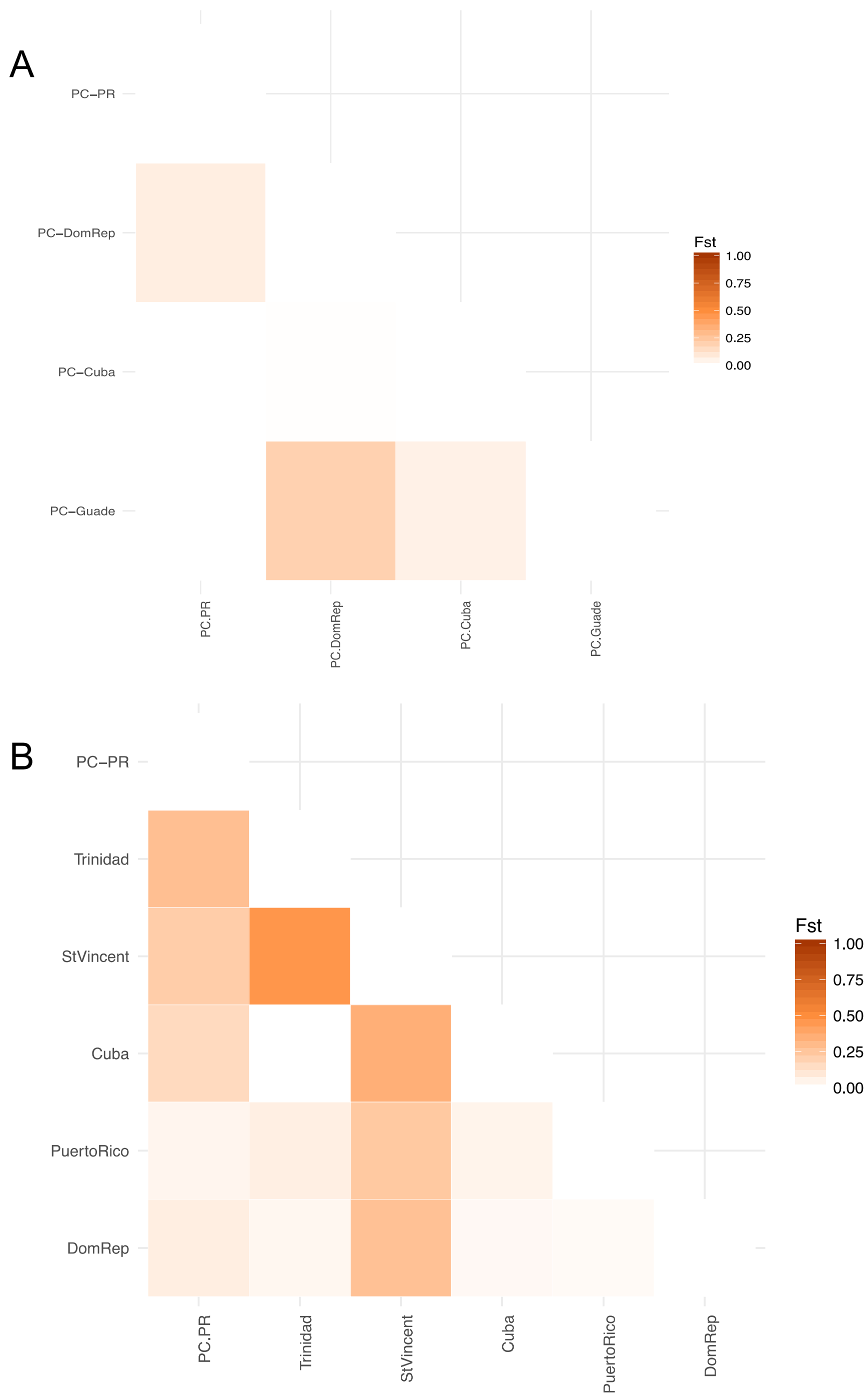


Figure 29. Fst genetic distances from HVR-I sequence data from ancient (A) and modern (B) Caribbean populations. Heatmap matrix of raw pairwise *Fst* values. Darker colors represent increased genetic differentiation. PC = pre-contact.

The last pairwise *Fst* matrix compared PC-PR to extant Caribbean populations (only Native American mtDNA lineages). This analysis found the lowest differentiation values between PC-PR and modern Puerto Ricans (Figure 29). However, the two populations were still significantly different in the permutation test (Table SM12). *Fst* pairwise matrices calculated with Caribbean population data are restricted to the HVR-I region, therefore similarity between haplotypes may be inflated due to loss of the resolution provided by additional HVR-II and control region substitutions. Exact tests conducted with HVR-I data found significant differences in haplotype distribution between PC-PR and most pre and post contact Caribbean populations (Native American lineages only; $p<0.05$) (Table SM13). Since this test is based on differences in haplotype frequencies, the results may reflect genetic differences created by the high rates of unique lineages found in each island group (see section below on *HVR-I median joining networks*).

#### 4.4.4 *Genetic differentiation over time*

To detect changes in mtDNA diversity over time within PC-PR, the twenty-one radiocarbon dated samples were divided into three temporal clusters of coeval radiocarbon dates: (1) 491-671 calAD with n=6 individuals, (2) 824-1039 calAD with n=9, and (3) 1095-1270 calAD with n=6 (Figure SM1). Figure 30 represents the distribution of mtDNA haplogroups over the three temporal groups as stacked bar charts. Lineage C1b2 is present in high frequencies in all time periods. No statistically significant differences in parent haplogroup (A2, C1, D1) distribution were found when examined with Fisher's Exact Test (Table SM14). But, simulations indicated that each

comparison had less than 30% statistical power, which suggests that sample sizes per cluster may be too low to allow for detection of small differences. Analysis of pairwise *Fst* values was repeated including the three PC-PR temporal subgroups and comparative populations (Table SM15). *Fst* distances were extremely low between temporal groups, and the null hypothesis of no significant difference was not rejected. Mantel tests comparing genetic versus temporal distance did not find a statistically significant correlation between genetic and temporal distance across the three groups ($z=107.8$, $p=0.111$) or between either group and modern Puerto Ricans ($z=84.5$, $p=0.168$).

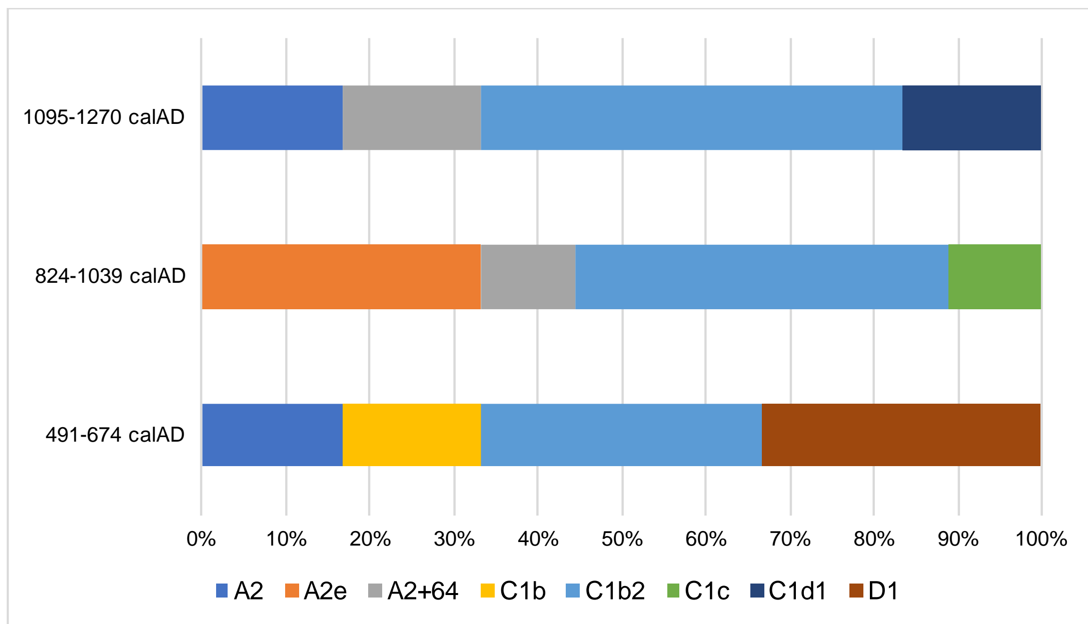


Figure 30. Distribution of mtDNA sub-haplogroups over time across all archaeological sites. Temporal groups identified through hierarchical clustering of median calAD radiocarbon dates for 21samples. Clusters are: (1) 491-671 median calAD with n=6 individuals, (2) 824-1039 calAD with n=9, and (3) 1095-1270 calAD with n=6.

#### 4.4.5 *Complete mtDNA median joining networks*

An initial round of analysis created median joining (MJ) networks comparing complete mtDNA sequences from PC-PR to all populations in the comparative dataset. Only haplogroups A2, C1 and D1 were considered. It was observed that PC-PR did not share haplotypes with any population except for extant Puerto Ricans. Therefore, network analysis was repeated excluding all other groups to clarify the relationship between the pre and post contact island populations.

The complete mtDNA MJ network for haplogroup A2 includes 8 ancient and 45 modern individuals (Figure 31). It is a diverse network, characterized by high frequencies of unique haplotypes (33 of 53 total sequences) and the lack of a central founder node. This topology suggests that most A2 lineages did not evolve in-situ but instead arrived to the island as derived haplotypes. The network suggests there was limited continuity of mitochondrial lineages over time. Only one pre-contact A2 lineage was shared with modern Puerto Ricans. In the ancient population, it is represented by individual PI420 from Paso del Indio. It is differentiated from other A2 types by polymorphisms 179C, 385G, 9947A and 16218T. The HVR-I motif for this lineage has also been observed among admixed Cubans and Puerto Ricans (Mendizabal et al. 2008, Vilar et al. 2014).

The complete mtDNA MJ network for haplogroup C1 includes 43 individuals, 15 from PC-PR and 28 modern Puerto Ricans (Figure 31). Network topology is characterized by several unique haplotypes and one large node composed of 18 identical sequences shared between the pre and post contact populations. The central lineage is haplogroup C1b2, defined in PhyloTree by control region polymorphisms G263A, C4242T, G7013A, C9557T, and G12454A. In the present network, C1b2 exhibits a star-

like phylogenetic pattern, where a founding, high-frequency lineage occupies a central position among several unique or low-frequency derived haplotypes (Forster 2004, Bandelt et al. 1995). This pattern is consistent with a history of expansion for this lineage. Haplogroup C1 and derived haplotypes are found in high frequencies in most ancient and modern Caribbean populations sampled to date (see next section) (Benn-Torres et al. 2015, Vilar et al. 2014, Mendizabal et al. 2008, Mendisco et al. 2015, Bryc et al. 2015, Lalueza-Fox et al. 2003, Martínez-Cruzado 2010, Lalueza-Fox et al. 2001). Lastly, complete mtDNA networks were not constructed for haplogroup D1 because no complete comparative data is available for modern Puerto Rican D haplotypes.

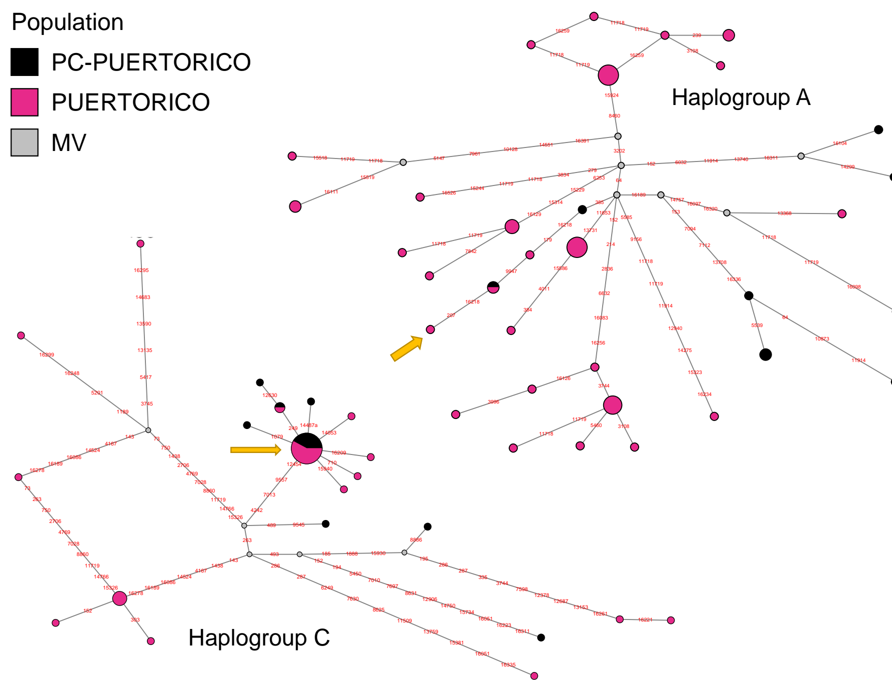


Figure 31. Complete mtDNA networks for haplogroups A2 and C1 including pre and post contact Puerto Rico populations. Mutations identified relative to the rCRS. Node size is proportional to the frequency of a given haplotype. Yellow arrows indicate lineages C1b2 and A2+16218 which were included in coalescent dating analysis. PC = pre-contact. MV = median vector

The coalescence date of haplotype branches in MJ networks can be estimated by equating the number of mutations differentiating each lineage with the substitution rate at the studied locus. This approach uses the molecular clock to date when a lineage arose or arrived in a population and its subsequent expansion (Forster 2004). Coalescence times were estimated for the entire C1b2 cluster and the A2+16218 branch containing sample PI420. Both are indicated by yellow arrows in Figure 31. The age of the C1b2 cluster was estimated as 1,294 ± 517 YBP. Radiocarbon dates obtained for PC-PR sequences within this cluster fall between 1167 – 943 calAD (813 to 1013 YBP) and therefore are within the standard deviation range of the genetic date. The coalescence estimate for the A2 branch was 13,046 ± 5,976. This estimate is inconsistent with previous research (see *Discussion*). Sample PI420 has been radiocarbon dated to 1270 calAD, or approximately 620 YBP.

4.4.6 *HVR-I median joining networks*

A second round of MJ network analysis was conducted comparing PC-PR haplotypes to pre and post contact Caribbean populations. Since comparative data were not available for complete mtDNA genomes, analysis was restricted to HVR-I sequences. In the MJ network for haplogroup A2 (Figure 32) sequences from all populations make up the central node. Topology mirrors that of the complete mtDNA network although diversity is reduced and some branches which appeared as distinct in the previous analysis are now collapsed. 53 of the 139 HVR-I sequences are unique haplotypes. Ancient populations share HVR-I haplotypes with several modern groups. For instance, lineage A2+16218 and derived types are observed in PC-PR, PC-Guadeloupe and Cuba

(Mendisco et al. 2015, Mendizabal et al. 2008). A second lineage, distinguished by polymorphism 16336A and defined as A2e in Phylotree is shared between Cubans and PC-PR. In addition, several low-frequency haplotypes observed in modern populations derive from founder nodes that include pre-contact individuals. Despite these instances of lineage sharing, many low frequency haplotypes are island specific. Lineages from the Trinidad FPC, for example, form a single clade. This overall topology is not consistent with a single A2 Caribbean founder but instead suggests multiple independent introductions with subsequent expansion of some of the derived lineages. It may also suggest some differentiation in A2 haplotypes between the Greater and Lesser Antilles, although more data from pre-contact Lesser Antillean groups would be needed to assess this thoroughly.

In contrast, the HVR-I network for haplogroup C1 shows a clear star-like pattern suggestive of in-situ expansion. The C1 founder shared across all Caribbean populations is defined by HVR-I polymorphisms 16223T, 16298C, 16325C, 16327T. As with haplogroup A2, some lineages which appeared to be distinct in the complete mtDNA genome analysis are now collapsed, largely into the central node. Most low-frequency derived haplotypes are separated from the founder by one mutation and are present in just one island population. A derived branch defined by a polymorphism 16298C only contains pre-contact haplotypes. PC-Guadeloupe and PC-PR individuals in this cluster have coding region variants which place them within haplogroups C1d and C1c, but these variants are not considered in the network because they were not typed for the PC-DR and Lesser Antillean samples. This network is composed of 36 unique haplotypes distributed across 100 individuals.

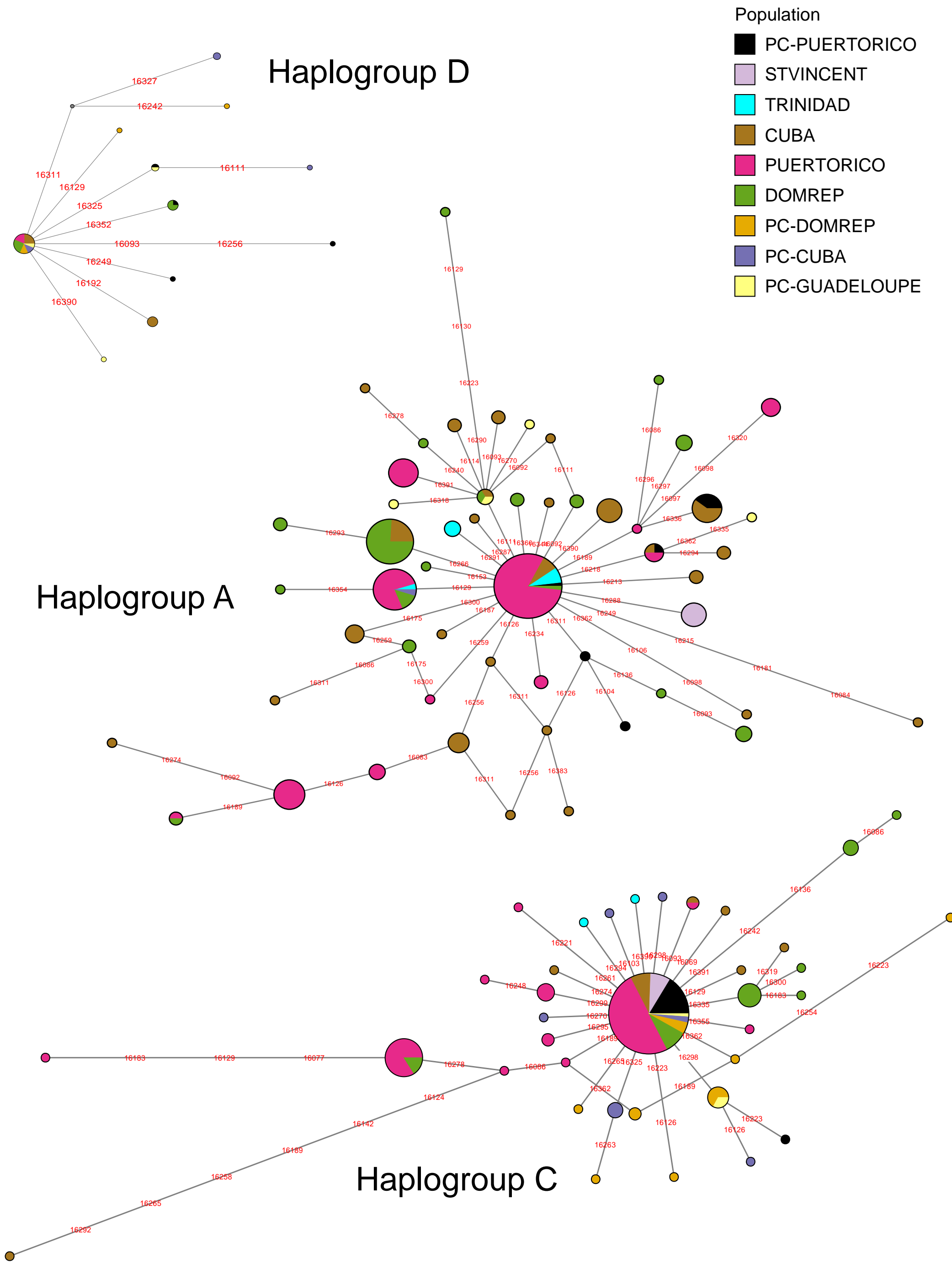


Figure 32. HVR-I networks for haplogroups A2, C1 and D1 including pre and post contact Caribbean populations. Networks are restricted to positions 16056 – 16391 which were sequenced across all datasets. Mutations identified relative to the rCRS. Node size is proportional to the frequency of a given haplotype. PC = pre-contact

Haplogroup D1 is present at low frequencies across all Caribbean populations. However, the HVR-I network indicates there is a large diversity of unique D1 lineages (21 individuals distributed into 11 haplotypes). This topology is not consistent with in-situ population expansion but instead suggests that derived D1 haplotypes arrived independently into the Antilles. PC-PR individuals carried four unique D1 haplotypes separated from the central node by one or two mutations. Two of these haplotypes are also observed in the Dominican Republic (Nieves Colón, Díaz-Zabala, and Martínez-Cruzado 2012) and in PC-Guadeloupe (Mendisco et al. 2015). Complete mtDNA sequences for these lineages are not present in the latest PhyloTree classification which suggests they have not yet been sampled in modern populations. Lastly, it must be noted that inferences from comparative analysis of HVR-I data were restricted to commonly sequenced regions across multiple genetic datasets. In some instances, haplotypes distinguished by coding region variants were grouped as equivalent taxa in the network estimation. This may have caused divergent clades to appear identical, and overestimated the amount of lineage sharing between ancient and modern groups, therefore underestimating the uniqueness of each island's mtDNA pool.

4.4.7 *Whole genome enriched samples: Preservation, enrichment success and sample authenticity*

Shotgun sequencing of thirteen aDNA libraries from the PC-PR sample revealed poor preservation of autosomal DNA. Average endogenous content was 0.45% for all libraries. Only two samples yielded over 1% endogenous content (Table SM6). Eight aDNA libraries were subjected to whole genome enrichment with the aim of recovering

autosomal genotypes. Only one of the enriched samples, PI 420, had high endogenous DNA content after WISC capture (10%). 1,688,415 unique reads mapping to the nuclear genome were recovered for PI420 after WISC capture but genomewide read depth was extremely low (0.035X). In contrast, 83% of the mitochondrial genome was recovered with an average read depth of 10X. These results are indicative of the differential preservation of mtDNA versus autosomal DNA due to the higher quantities of mitochondria in eukaryotic cells (Giles et al. 1980).

A second sample, PI 413, had 2.3% endogenous content after capture, but mtDNA read analysis revealed a non-Native American haplogroup assignment. In addition, deamination plots indicated this sample did not have authentic aDNA damage patterns (Figure SM4). Given that this pattern was also similar in the mtDNA enriched library (Table SM3 and SM5), it seems likely contamination was introduced to the sample extract or library at some point before enrichment. WISC enriched samples are captured independently (not pooled) and no other samples had evidence of potential exogenous contamination. Deamination patterns for sample PI420 were consistent with expectations for authentic ancient DNA (Figure 33) (Dabney, Meyer, and Pääbo 2013, Briggs et al. 2007).

4.4.8 *Principal components analysis and ADMIXTURE estimation of autosomal data*

PCA was conducted to determine whether sample PI420 showed closed affinity to specific reference populations. SNP genotypes at 11,742 autosomal sites intersected with the HGDP reference panel were used as input for multivariate analysis. The panel contains 51 global populations including five Native American groups: Karitiana, Maya,

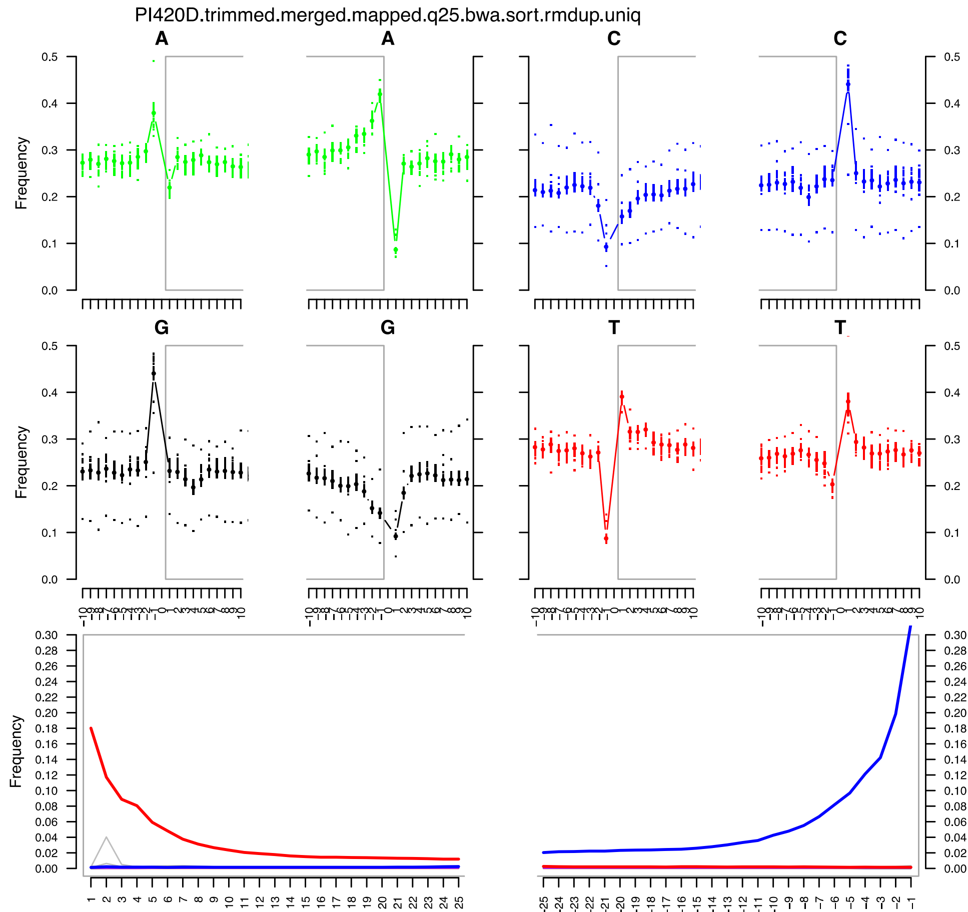


Figure 33. Damage plots and fragment misincorporation patterns for autosomal reads recovered from sample PI420 after WISC enrichment. Plot generated with mapDamage.

Pima, Surui and Piapoco (Colombian) (Rosenberg 2006, Rosenberg et al. 2002). PI420 fell far from Native American populations in PC space (Figure 34), clustering close to Oceanian aboriginal groups in PC 1 vs 2. This pattern is recapitulated at higher components as well (Figure SM5A). To examine whether this was the artifact of post-mortem damage, SNPs with C to T or A to G transitions, which are likely to arise due to deamination, were removed from the dataset. A second PCA was conducted on the remaining 10,754 SNPs but the position of PI420 in PC space did not change (Figure

SM5B-C). Lastly, the effects of modern contamination were evaluated by repeating the PCA on autosomal reads filtered for PMD scores >0 and >3, keeping 5,492 and 3,249 SNPs, respectively, after each filtering run. PCA were repeated a third time on the damage filtered datasets. Even though resolution is reduced due to the lower density of input data, the position of sample PI420 on the plot remained the same (Figure SM5D-G).

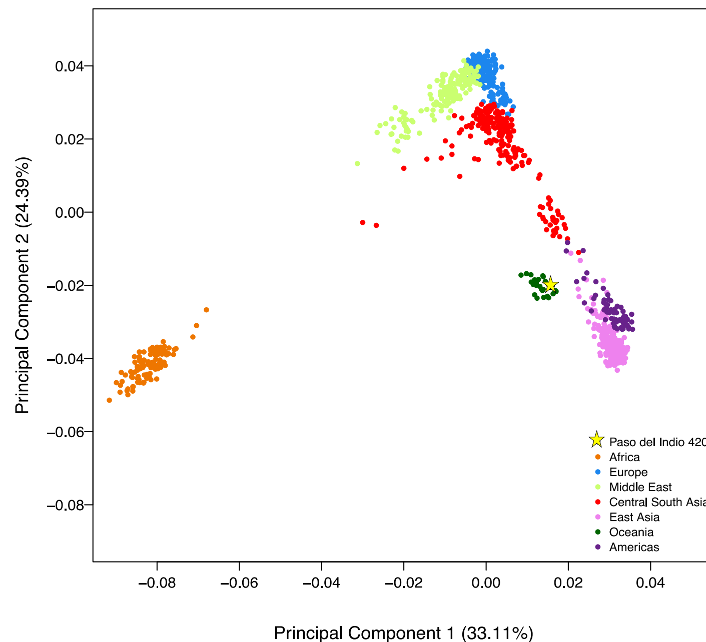


Figure 34. PCA conducted with 11,742 autosomal SNPs comparing PI420 to 51 HGDP populations (PC 1 vs 2). Sample PI420 shown as a yellow star.

To explore the genetic ancestry of sample PI420 further, unsupervised clustering analysis was performed in ADMIXTURE including all HGDP populations. Clusters from K=2 to K=6 were explored, and the lowest cross-validation error was observed at K=*6*. Stacked barplots shown in Figure 35 clearly indicate that sample PI420 consistently shows a three-way ancestry pattern from K=3 onwards. At K=6, this pattern is

characterized by a large dark blue component shared with Native American populations and smaller yellow and light blue components shared with Eurasian and African populations, respectively. The estimated ancestry proportions for PI420 were 74.8% Native American, 10.3% African and 13.9% Eurasian, with the remaining ancestry made up of >1% from other sources.

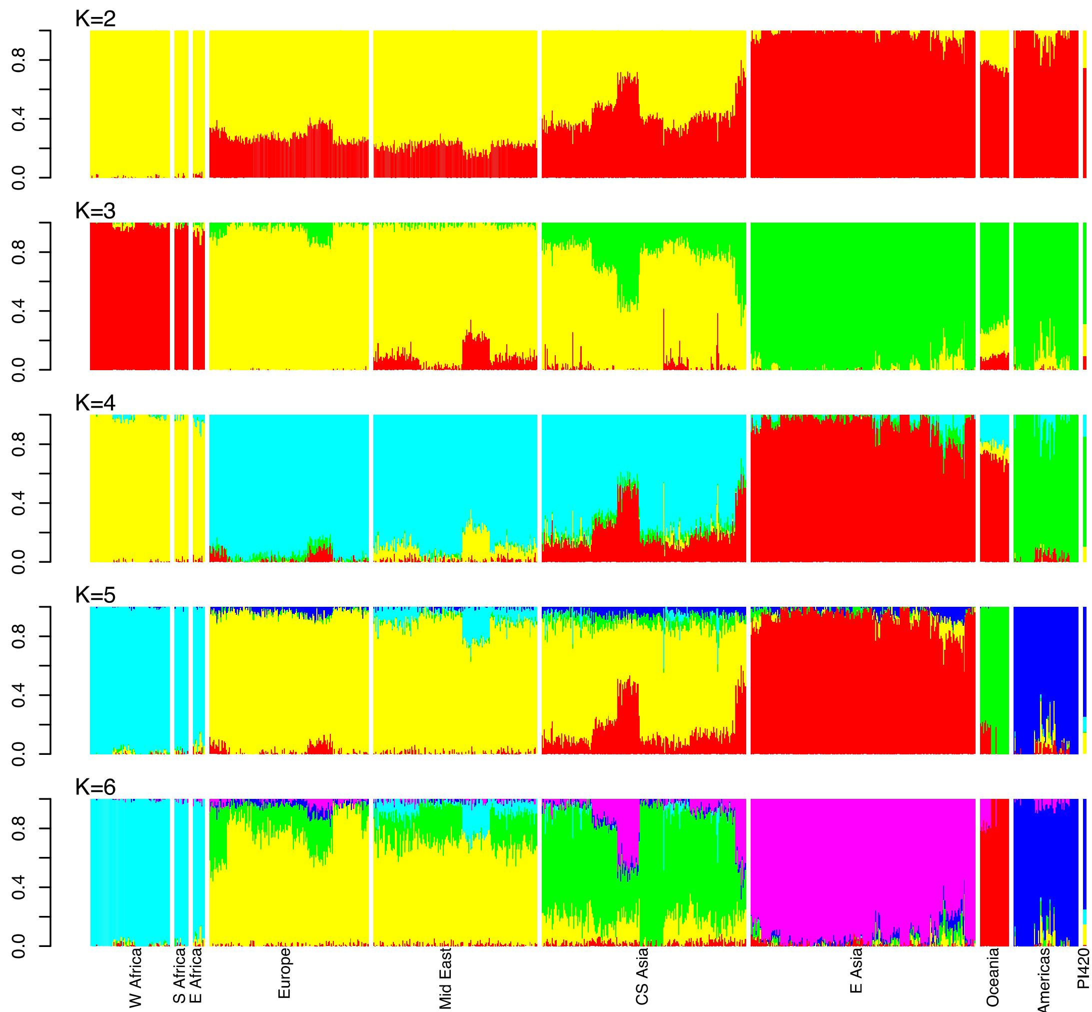


Figure 35. ADMIXTURE analysis K=2 to K=6 with 11,090 autosomal SNPs comparing PI420 to HGDP populations. SNPs were thinned for linkage disequilibrium. PI420 sample is the last bar on the right.

## 4.5 Discussion

4.5.1 *Ancient DNA preservation in Puerto Rican archaeological sites*

Moderate to high coverage mtDNA genomes were recovered from 27 of the sixty ancient Puerto Rican remains. This represents a success rate of approximately 45%. Differential rates of aDNA preservation were observed between samples collected at Paso del Indio, Punta Candelero and Tibes. This suggests that in the tropics, site-specific processes may play a larger part in aDNA decay than island or region-wide environmental conditions. Samples from Tibes had the poorest genomic preservation. This finding is consistent with previous reports of poor collagen and organic preservation in skeletal remains from Tibes compared to Paso del Indio, Punta Candelero and other pre-contact sites in Puerto Rico (Pestle and Colvard 2012). Future research with biological remains from this site may benefit from assessing the extent of endogenous aDNA in other dense tissues, such as the petrous portion of the temporal bone, or in mineralized dental calculus (Hansen et al. 2017, Pinhasi et al. 2015, Ozga et al. 2016). In this dataset, a slight positive correlation was observed between sample age (measured in calibrated radiocarbon dates) and mtDNA recovery. This suggests that samples with younger ages tend to yield higher endogenous content after mitochondrial capture. However, the correlation test did not control for inter-site effects and could be influenced by other factors such as capture efficiency or sequencing effort. Additionally, because it is known that enrichment can bias sequence composition (Ávila-Arcos et al. 2015), this relationship should be further tested by large scale shotgun sequencing of radiocarbon dated remains from the selected sites. This approach would hold great potential for better

characterizing the relationship between thermal age and rates of DNA fragmentation and degradation in tropical contexts (Hofreiter et al. 2015).

### 4.5.2 *Whole genome data*

Nuclear genome coverage was estimated in a subset of the remains before and after whole genome enrichment. This comparison demonstrated that autosomal DNA is preserved at extremely low rates in ancient Puerto Rican remains. Poor aDNA preservation has been observed in previous research conducted with tropical remains and was not unexpected for this dataset (Schroeder et al. 2015, Lalueza-Fox et al. 2003, Gutierrez-Garcia et al. 2014). The results of shotgun sequencing highlighted the importance of target enrichment approaches for increasing chances for tropical aDNA recovery.

Autosomal genotypes produced from one individual provided inconclusive results due to low read depth, sparse genome coverage, excessive DNA degradation, and potentially, exogenous contamination. PCA plots indicated sample PI420 clustered close to aboriginal Oceanian populations in PC space (Figure 34). This result does not necessarily indicate that PI420 has a strong link with Oceania. Admixed individuals tend to fall on a clinal gradient between parental groups in PCA (Wollstein and Lao 2015, Ma and Amos 2012). Therefore, the position of the data point suggested PI420 could carry some genomic admixture with non-Native American populations. This was confirmed with the genomic ADMIXTURE barplots and global ancestry proportion estimates.

These results were unexpected because the radiocarbon date for PI420 precedes European contact (1270 calAD) and mtDNA data indicates this individual carried a

Native American haplogroup. It is possible that the signature of genomic admixture from multiple continental populations stems from sample contamination. However, WISC enriched sequence reads recovered for PI420 show the characteristic signs of authentic aDNA (increased deamination at 5’ and 3’ ends and small average fragment size). Further, WISC recovered mtDNA reads were estimated to be ~99% authentic which is consistent with recovery of endogenous aDNA (Table SM6, Figure 33). High rates of post-mortem damage could also explain skews in the data, but the observed pattern persists after excluding variants that could potentially arise from cytosine deamination. Another possibility is that the extremely low coverage and read depth of the recovered reads may have been insufficient to generate confident SNP and genotype calls (Parks and Lambert 2015). This last issue can be addressed through additional sequencing and application of newer methods developed for autosomal variant calling and duplicate read removal in aDNA datasets (Peltzer et al. 2016). Lastly, although unlikely given the occupation history of Paso del Indio, the possibility that there were errors in radiocarbon date estimation or that this individual is from an intrusive post-contact burial cannot be discarded. Given these results, additional analysis of autosomal reads recovered from the whole genome enrichment were discontinued and the results are not discussed further in the present section. A second effort to recover nuclear DNA from ancient Puerto Rican remains is currently ongoing.

#### 4.5.3 *MtDNA diversity of pre-contact Puerto Rican communities*

Analysis of mtDNA diversity identified high frequencies of Native American haplogroups A2 and C1, and low frequencies of haplogroup D1 in the ancient population.

Similar haplogroup distributions have been previously reported in several ancient and modern Caribbean populations from across the Greater and Lesser Antilles, including modern Puerto Ricans of Native American ancestry. This indicates that the PC-PR population follows a general Caribbean wide pattern of mtDNA haplogroup diversity and distribution (Benn-Torres et al. 2015, Mendizabal et al. 2008, Mendisco et al. 2015, Martínez-Cruzado 2010, Martínez-Cruzado et al. 2005, Vilar et al. 2014, Nieves Colón, Díaz-Zabala, and Martínez-Cruzado 2012, Benn-Torres, Kittles, and Stone 2007, Bryc et al. 2015).

Genetic relationships at an intra-island scale were evaluated by testing for significant differences in mtDNA diversity between coeval individuals from the three studied sites. Archaeological evidence indicates there was a growing trend towards cultural differentiation and regionalization across Caribbean chiefdoms during the Late Ceramic Age (CA). Within Puerto Rico, this is visible through the emergence of distinctive material culture, settlement patterns and ceramic traditions that divided communities along an East-West divide after the year A.D. 600 (Rouse 1992, Curet and Stringer 2010, Curet 2005, Curet, Torres, and Rodríguez 2004). In this study, I find no evidence for genetic structure or significant differences in mtDNA haplogroup composition between communities at Punta Candelero, Paso del Indio and Tibes. Although individuals compared from each population are not strictly coeval, temporal cluster analysis also finds no evidence for the emergence of temporal structure or differences in mtDNA haplogroup composition over time in this population. This finding contrasts with the analysis conducted by (Martínez-Cruzado et al. 2005) who found a strong geographic gradient in the distribution of C1 lineages among modern Puerto

Ricans, but is consistent with subsequent findings by (Vilar et al. 2014, Gravel et al. 2013). This suggests that the differentiation patterns observed by (Martínez-Cruzado 2010) could stem from more recent demographic processes. A lack of strong genetic differentiation between PC-PR communities indicates that intra-island cultural diversification during the late CA was not accompanied by genetic isolation or restrictions of female-mediated gene flow.

Interestingly, palaeodemographic reconstruction using skeletal remains from Punta Candelero suggests this site had large numbers of young and middle aged adult males, a pattern that is consistent with an immigrant population (Curet 2005). In contrast, the demographic profile and archaeological evidence for Paso del Indio and Tibes indicate both sites had long term and stable occupations, despite some periods of abandonment (Curet 2005, Walker 2005, Rodríguez López 1991, Curet and Stringer 2010). I find no evidence for increased genetic diversity at Punta Candelero relative to Paso del Indio, and the small number of samples recovered from Tibes precludes more extensive efforts at inter-site comparisons. Although additional sampling from each of these populations is currently underway, the poor aDNA preservation at Tibes may not allow for systematic reconstruction of the genetic diversity of this site's population.

MtDNA diversity in PC-PR is low compared to most continental Native American populations, except for the Surui, Karitiana and Pima (Table 2). Low genetic diversity has previously been reported for Amazonian populations from Eastern South America due to small historical effective population sizes and repeated genetic bottlenecks (Lewis et al. 2007, Hunley, Gwin, and Liberman 2016). Similarly, low haplotype diversity in PC-PR may stem from genetic drift and a strong founder effect

during the original peopling of the region, as shown by the high frequency of C1b2 lineages. Founder effects could have occurred during the initial island settlement or during the subsequent Ceramic Age expansions. Previous studies of mtDNA and autosomal loci among modern Puerto Ricans noted a pattern consistent with strong effects of drift and at least one population bottleneck in the demographic history of Caribbean Native American groups (Gravel et al. 2013, Martínez-Cruzado et al. 2005). The aDNA mitochondrial data analyzed here supports a scenario where a reduced number of mtDNA haplotypes became isolated from their ancestral population after settling in Puerto Rico. This led to a reduction in effective population size (*Ne*), loss of ancestral diversity and increased susceptibility to genetic drift (Hedrick 2011). Subsequent expansion of one of the founder lineages, C1b2, then created the star-like pattern visible in the phylogenetic network of Puerto Rican C1 lineages (Bandelt et al. 1995, Forster 2004).

Evidence for a dramatic effect of genetic drift on mtDNA diversity in the Greater Antilles has been previously described in studies of pre and post-contact Cuban and Dominican populations (Lalueza-Fox et al. 2001, Lalueza-Fox et al. 2003, Mendizabal et al. 2008). In contrast, communities from PC-Guadeloupe have high mtDNA diversity, despite the low genetic diversity that is observed today among the Trinidadian FPC and Vicentian Garifuna groups of neighboring Lesser Antilles (Benn-Torres et al. 2015, Mendisco et al. 2015). In theory, a pattern of increased diversity in the Lesser versus the Greater Antilles would be consistent with archeological hypotheses of a west to east stepping stone migration where South American populations entered the archipelago via Trinidad and rapidly moved upwards, reaching the Greater Antilles in less than a

thousand years (Keegan 1995, Rouse 1986). However, this model would predict that genetic diversity would be higher, not lower in PC-PR than in PC-Cuba or PC-DomRep.

Inferences regarding the direction of pre-contact Caribbean migrations cannot be drawn exclusively from interpretation of summary statistics, especially given the widely different dates and small sample sizes available from comparative populations (Mendisco et al. 2015, Lalueza-Fox et al. 2001, Lalueza-Fox et al. 2003). Future application of more complex approaches such as Bayesian modeling and testing of explicit migration scenarios while considering time-stamped sequences will allow for more robust testing of inter-island migration scenarios.

4.5.4 *MtDNA haplogroups and the continental origins of Caribbean indigenous groups*

Statistical analyses of genetic differentiation values find similarities between the mtDNA gene pool of PC-PR and extant indigenous populations from northern South America and Amazonia, including groups from Venezuela, Guyana, Brazil and Colombia. This finding is consistent with previous genetic research with mtDNA and nuclear loci (Gravel et al. 2013, Moreno-Estrada et al. 2013, Martínez-Cruzado 2010, Martínez-Cruzado et al. 2005, Vilar et al. 2014), and with archaeological evidence placing the ultimate origin of Ceramic Age Caribbean population among Arawak speaking cultures of the Orinoco region (Chanlatte Baik, 2003; Siegel, 2005, Rouse 1986; Siegel 2005). But, the results of the exact test do not reject a possible contribution from Mesoamerica to the ancestry of late Ceramic Age PC-PR populations.

I find no evidence of genetic affinity between the Chippewa and Pima and PC-PR. This finding is consistent with lack of archaeological evidence for any contribution from

North American populations to the cultural ancestry of Caribbean indigenous groups (Pantel 2003). However, the Chippewa and Pima reside in southwest and northeast North America. Julian Steward, in his Handbook of South American Indians, proposed the Florida peninsula, as a potential source area for Caribbean groups (Steward 1947), and recent genetic data has suggested potential movements southwards into the Antilles from the eastern north American coast (O'Rourke and Raff 2010). Therefore, incorporation of genetic data from extant or ancient groups native to the Southeastern United States could prove to be a more informative comparison. Unfortunately, such data is not currently available in public catalogs of human genetic variation.

Interestingly, *Fst* permutation tests rejected the null hypothesis of no genetic differentiation between PC-PR and every continental population in the comparative dataset. However, no significant differentiation was observed between PC-PR and pre-contact communities from Cuba and the Guadeloupe archipelago. These results suggest that local evolutionary processes such as drift may have led to differentiation of the Caribbean mtDNA gene pool from that of continental indigenous groups. The only exception to this pattern seems to be the PC-DomRep population which is distinguished from other ancient Caribbean groups by an absence of A2 haplogroups (Lalueza-Fox et al. 2001). However, the HVR-1 median network indicates that modern Dominicans carry large frequencies of A2 lineages and some of these are identical or derived from haplotypes found in other ancient Caribbean populations. This suggests that at least some of the extant A2 lineages may have a pre-contact origin. Testing this prediction would require additional sampling of human skeletal remains collected from Ceramic Age sites in Dominican Republic or Haiti.

The most frequently observed mtDNA haplogroup in PC-PR was C1b2, which made up 44% of all ancient lineages. C1b2 is derived from haplogroup C1b, which is one of the nine New World founding lineages (Tamm et al. 2007). Although C1b is found across the Americas, it is seen in highest frequencies among South American native groups such as the Yanomami (Williams, Chagnon, and Spielman 2002, Perego et al. 2010). Coalescent analysis places C1b2's likely origin in South America approximately 2,000 YBP (Gomez-Carballa et al. 2015). It has been identified through restriction fragment length polymorphism (RFLP) assays and control region sequencing in admixed Brazilians and Uruguayans, and among Brazilian and Venezuelan Amazonian tribes such as the Yanomami and the Kraho (Torroni et al. 1993, Cardena et al. 2013, Sans, Mones, et al. 2015).

C1b2 is also the most common C1 lineage observed in modern Puerto Ricans (Vilar et al. 2014, Martínez-Cruzado 2010). The large number of identical C1b2 haplotypes and its consistently high frequency over time provide the strongest continuous link between the pre and post contact population. This clade has a clear signature of population expansion in the complete mtDNA networks and its estimated arrival date of 1,294 ± 517 YBP falls within the Ceramic Age period. This estimate differs from that obtained by (Vilar et al. 2014) for the complete control region (647 ± 373 YBP) but is consistent with previous analyses conducted by (Martínez-Cruzado 2010) on HVR-I haplotypes (1,195 ± 690 YBP).

These findings are consistent with previous research (Vilar et al. 2014, Martínez-Cruzado 2010). Taken together these data suggest that the Puerto Rican C1b2 clade arrived from South America, possibly carried by Arawak migrants during the Ceramic

Age expansion into the Antilles. Unfortunately, control region polymorphisms typed in most other Caribbean populations do not provide sufficient resolution to distinguish C1b2 from the C1 founder lineage. However, the large frequency of C1 haplotypes observed across the region's ancient and modern peoples (Lalueza-Fox et al. 2001, Lalueza-Fox et al. 2003, Mendizabal et al. 2008) and the identification of C1b lineages in pre and post contact Lesser Antillean populations (Mendisco et al. 2015, Benn-Torres et al. 2015) support designating this clade as a characteristic motif of Ceramic Age Antillean populations (Vilar et al. 2014, Martínez-Cruzado 2010, Schurr 2010).

Two additional C1 haplogroups, C1c and C1d1, were also identified in PC-PR. Both are founding New World lineages and have a wide distribution across the Americas (Perego et al. 2010, Tamm et al. 2007). C1c was identified in individual T257 from Tibes. This haplogroup is distinguished by coding region polymorphisms G1888A and G15930A (PhyloTree Build 17). Although its present distribution is widespread, C1c is common among Mexican and Mexican American populations (Perego et al. 2010, Kumar et al. 2011). Mendisco et al. (2015) found one C1c haplotype among late Ceramic Age individuals from PC-Guadeloupe but this lineage differs from that carried by individual T257 by one mutation at site T16311C. Although, C1c lineages have also been observed in modern Puerto Ricans, none of these haplotypes match the T257 sequence.

Haplogroup C1d1 was observed in the PCPR sample in individual PI419, from Paso del Indio. C1d1 is distinguished by four polymorphisms: A16051G, C194T and G7697A (PhyloTree Build 17). Basal C1d lineages are common across Central and South America (Perego et al. 2010, Figueiro, Hidalgo, and Sans 2011). For instance, one C1d derived type, C1d3 was characterized from pre-contact remains in Uruguay (Sans,

Figueiro, et al. 2015). Other C1d variants have been seen in extant Mexican and Andean populations (Guardado-Estrada et al. 2009, Barbieri et al. 2011). In the Caribbean, C1d1 has been found in Cuba, Dominican Republic and the Trinidadian FPC (Mendizabal et al. 2008, Nieves Colón, Díaz-Zabala, and Martínez-Cruzado 2012, Benn-Torres et al. 2015). But none of these haplotypes match the one carried by individual PI419. Overall, these results suggest that C1c and C1d may have been introduced independently into the Antilles. The specific origin of the PC-PR C1c and C1d haplotypes within the continental Americas remains unclear.

Haplogroup A2 makes up 29% of the PC-PR mitochondrial pool and is represented by three major lineages: A2, A2+64 and A2e. Seven unique haplotypes are found in eight ancient individuals. Basal A2 lineages are among the nine New World founders and as such are widely distributed across the Americas. Nonetheless, North and Central American populations have higher frequencies and larger varieties of A2 haplotypes than other continental groups (Perego et al. 2010, Tamm et al. 2007, Gonzalez-Martin et al. 2015). All A2 haplotypes identified in PC-PR were distinct from lineages characterized in broad surveys of extant Mesoamerican mtDNA diversity; including studies conducted with admixed and indigenous peoples from Mexico, Panama and Guatemala (Kumar et al. 2011, Gonzalez-Martin et al. 2015, Söchtig et al. 2015, Mizuno et al. 2014, Perego et al. 2012, Gorostiza et al. 2012). For instance, four individuals in PC-PR carried derived lineages of haplogroup A2e also observed in admixed Cubans (Mendizabal et al. 2008). Although A2e is listed in a broad phylogeny of American haplotypes by (Achilli et al. 2008), its origin and geographical distribution is unknown. Difficulty in identifying links between Caribbean and mainland A2 haplotypes

does not necessarily exclude a Mesoamerican contribution to the peopling of Puerto Rico and the other Antilles. Genetic drift and the population bottlenecks caused by European contact could have led to loss of these lineages in extant mainland populations. A trend for genetic discontinuity between pre and post contact mtDNA diversity has previously been observed in large scale studies of ancient South American groups (Llamas et al. 2016).

Although A2 is the most common Native American haplogroup found in modern Puerto Ricans (Martínez-Cruzado et al. 2005, Vilar et al. 2014), only one derived haplotype was shared between pre and post contact populations: A2+16218. This haplotype was obserbed by (Martinez-Cruzado et al. 2001, Martínez-Cruzado 2010) based on RFLP and HVR-I sequencing among self-identified Native American descendant communities in Indiera Alta, Maricao, Puerto Rico. Subsequently, it has also been found in Cubans and Puerto Ricans and in one pre-contact sample from the site of Grande Anse in Guadeloupe (Mendizabal et al. 2008, Vilar et al. 2014, Mendisco et al. 2015).

Martínez-Cruzado (2010) inferred that despite its low frequency among contemporary Puerto Ricans, this lineage and its derived haplotypes may have arrived during the early pre-contact era, before the Ceramic Age expansions. This inference was based on its high diversity despite the lack of a population expansion signature. Attempts to test this hypothesis by dating A2+16218's arrival to Puerto Rico were unsuccessful. The coalescent estimate of 13,046 ± 5,976 pre-dates the archaeological evidence for human arrival into the Antilles and is not consistent with previous research (Martínez-Cruzado 2010). Coalescence estimates are highly influenced by the amount of descendant

lineages within a branch and how well the founder lineage confirms to a star-like pattern (Forster 2004). Therefore, the early date and large standard errors may be attributable to a small number of haplotypes included and the lack of a clear expansion signature. Nonetheless, the presence of this haplotype at Paso del Indio and across the Antilles supports a pre-contact arrival and suggests this lineage may have had a wide inter-island distribution. Unfortunately, the continental origin of A2+16218 is currently unknown and it is not included in the most recent PhyloTree build.

The overall topology of the complete mtDNA network for Puerto Rican A2 haplogroups is not consistent with a star-like phylogeny. Instead, it suggests that A2 haplotypes arrived to the island via multiple independent introductions with subsequent expansion of several derived lineages. This is consistent with inferences drawn from A2 networks estimated in previous research with HVR-1 sequences from Puerto Rico (Vilar et al. 2014, Martínez-Cruzado 2010). Based on the observation that A2 haplotypes in Cuba are more numerous and diverse than those in Puerto Rico, Vilar et al. (Vilar et al. 2014) proposed that two lineages found among modern Puerto Ricans, A2k and A2*, may have originated in Cuba and then moved eastwards. This scenario would lend support for west to east migration pattern throughout the Antilles. The HVR-1 networks constructed in the present research provide some support for this hypothesis. In general, Cuban and Dominican A2 lineages are more diverse than Puerto Rico A2 haplotypes, many of which are derived from Dominican and Cuban parental nodes. However, the aDNA data is not consistent with this scenario. Lineages A2k and A2* were not identified in PC-PR, and pre-contact populations from Cuba and the Dominican Republic have very few A2 haplotypes (Lalueza-Fox et al. 2001, Lalueza-Fox et al. 2003). More

sequence data from ancient Greater Antilles communities could further illuminate this issue. Lastly, despite some instances of lineage sharing across islands, the HVR-1 network for haplogroup A2 has multiple low-frequency and geographically restricted lineages which are only found on specific islands. This is especially notable for the Trinidad FPC population who did not share a single A2 haplotype with any other ancient or modern Caribbean group (Benn-Torres et al. 2015). This observation suggests that there may have been some degree of genetic isolation and differentiation between island communities during antiquity.

Although haplogroup D1 was only identified in four PC-PR samples, each complete mtDNA haplotype was unique. The haplotypes had several unique mutations differentiating them from the basal D1 lineage (defined by C2092T and T16325C). None of them were listed in PhyloTree and therefore could not be classified beyond the basal D1 type. This suggests that complete mtDNA lineages matching these haplotypes have not yet been sampled in extant populations. D1 is a New World founding haplogroup and therefore has a Pan-American distribution. However, basal and derived D1haplotypes are most commonly observed in South American indigenous populations (Perego et al. 2010, Tamm et al. 2007). Lalueza-Fox et al. (2001) found high frequencies of haplogroups D1 and C1 in pre-contact individuals from the site of La Caleta, Dominican Republic. Based on this distribution, they proposed a South American origin for PC-DomRep populations, and hypothesized that these were the direct descendants of Arawak migrants. Subsequently, Bodner et al. (2012) speculated that some of the PC-DomRep D1 lineages could correspond to haplogroup D1j which is geographically restricted to the Southern Cone.

None of the PC-PR D1 lineages matched the HVR-1 motifs identified in PC-DomRep (Lalueza-Fox et al. 2001) or the coding region polymorphisms that define D1j. However, this finding does not necessarily reject a South American origin for Caribbean D1 haplotypes. The D1 lineage carried by sample PCE54 from Punta Candelero matched another Ceramic Age individual from the site of Grotte Cadet 2, Marie Galante, Guadeloupe (Mendisco et al. 2015). This haplotype lacks polymorphism T16325C and was previously reported by (Alves-Silva et al. 2000) in a single individual from northern Brazil. Additionally, the lineage carried by individual T251 from Tibes is an incomplete match to an unclassified D1 haplotype identified by (Lee and Merriwether 2015) among the Yekuana tribe of the Venezuelan Amazon. Although the two lineages are not identical, they share many rare mutations and thus may derive from the same common ancestor.

Based on network analysis of modern Puerto Rican HVR-I sequences Martínez-Cruzado (2010) predicted that haplogroup D1should be present at low frequencies in PC-PR. Results from the present work support this prediction, but the lack of a star-like phylogeny prevents robust estimation of the arrival date of these lineages into the Caribbean. Interestingly, no lineage sharing was observed for D1 HVR-1 lineages between pre and post contact Puerto Rico, but a match was found between sample T251 from Tibes and three Dominican individuals. The present distribution of Caribbean D1 lineages suggests that this haplogroup was rare during antiquity. However, the diversity of PC-PR D1 haplotypes and the fact that they are shared with multiple ancient and modern populations suggests that haplogroup D1 frequencies in the Caribbean today may

be skewed due to genetic drift or to the extreme lineage loss caused by post-contact population bottlenecks (see next section).

### 4.5.5 *MtDNA diversity across the pre and post contact divide*

When comparing HVR-1 mtDNA variation in PC-PR with that of extant Caribbean groups (Native American lineages only), the lowest genetic differentiation was observed with modern Puerto Ricans. Although this suggests there is a link between pre and post contact island populations, the complete mtDNA exact tests, *Fst* matrices and MDS plots detect significant differences between the two groups. This finding stems from the different haplogroup composition of ancient and modern Puerto Rican populations. As discussed above, only two mtDNA haplotypes were shared across the two samples. Several reasons could account for this finding.

*First*, the poor aDNA preservation of the remains restricted the pre-contact sample to just 27 individuals. Even though this is the largest sample of aDNA data generated from ancient Caribbean populations to date, it is possible that the full extent of mtDNA diversity present in PC-PR is not represented in this dataset. Ongoing research including sixty additional remains from these three sites will at least partially account for this problem. Given the success rates obtained during this project (45%) I estimate 27 additional mitochondrial genomes to be recovered from this additional sampling effort.

*Second*, if some of the mtDNA lineages observed in the modern population arrived after the last known occupation of the sites included in this study (i.e. 1300 A.D.), they would be undetectable. In the modern population these haplotypes would appear as isolated, derived types in network analysis and would not present signatures of

population expansion (Martínez-Cruzado 2010). Unfortunately, Native American lineages that arrived after European contact would also share these characteristics.

*Third*, genetic drift can lead to stochastic loss of genetic diversity over time, especially in isolated populations with high rates of inbreeding, skewed sex ratios, or few breeding pairs (Charlesworth 2009). This effect is even more pronounced for mtDNA which, due to its haploid nature and uniparental inheritance, has about a fourth of the effective population size (*Ne*) of nuclear loci (Hedrick 2011). For example, Helgason et al. (2009) found that drift was the most likely explanation for large differences in mtDNA lineage composition between modern Icelanders and their Medieval ancestors. This pattern could not be explained by migration, but instead was consistent with stochastic loss of ancient haplotypes due to repeated genetic bottlenecks and reductions in female *Ne.* As discussed above, indigenous populations in the Caribbean likely experienced a strong reduction in *Ne* due to founder effect during the settlement of the islands. This would have produced a genetic pool characterized by few mtDNA haplotypes and thus susceptible to stochastic changes in lineage frequencies or loss of low frequency haplotypes. Therefore, even without the large scale demographic changes that occurred after contact (see below), it is possible and even likely that several mitochondrial haplotypes found in PC-PR would not be passed on to modern Puerto Rican populations.

*Fourth,* there is extensive documentary evidence of a sharp reduction in population size across the Caribbean islands during the conquest period. Historians have estimated that shortly before European contact the population of Puerto Rico was anywhere between 30,000 to 70,000 people, with some estimates as high as 1 million (Cobley 1994, Anderson-Córdova 2005). By the census of 1530, however, the total native

population was reported to be about 1,543 individuals (Anderson-Córdova 2005). Among these, some were also Native Americans forcibly relocated from other islands or from the continental Circum-Caribbean basin (Anderson-Córdova 2005, Wilson 2007). Colonial era census likely under-represented the true number of indigenous people living in Puerto Rico and other Caribbean colonies throughout the Spanish occupation, therefore these estimates must be interpreted with caution (Benn-Torres 2014). Estimates of long term historical *Ne* have been gleaned from autosomal genome fragments found in modern Puerto Ricans from the 1000 Genomes Panel. These suggest that the Native American ancestors of the modern population had an effective breeding population size of just 1,922 individuals; 32 times smaller than the estimated size of ancient Mexican populations (Gravel et al. 2013). Thus, modern Puerto Ricans likely carry a subset of the mtDNA diversity that was present in the ancient pre-contact population, along with several mtDNA lineages introduced during the colonial era.

My findings do not support a scenario of complete population replacement or extinction of indigenous Caribbean populations. In a database of 735 complete mitochondrial genomes from throughout the pre and post-contact Americas, the only identical matches to PC-PR haplotypes where found among modern Puerto Ricans. Although this may partly be an artifact of under sampling among continental Native American populations (1000 Genomes Project et al. 2015, Bustamante, De La Vega, and Burchard 2011) it also points to the large role that indigenous groups have played in shaping the genetic diversity of admixed Caribbean peoples.

## 4.6 Conclusion

This research characterizes the genetic diversity of pre-contact Puerto Rican groups and tests hypotheses about their relationships and interaction with other Native American and Caribbean populations. Complete mtDNA genomes recovered from the ancient inhabitants of Punta Candelero, Paso del Indio and Tibes suggest they shared a common origin and may have engaged in continuous genetic exchange. This indicates that a process of cultural diversification observed in the archaeological record during the late Ceramic Age was not accompanied by genetic isolation or reduced female-mediated gene flow between local communities. I also identify a shared genetic component among inter-island pre-contact groups, mainly represented by shared C1 haplotypes and similar mtDNA haplogroup frequencies. This is in line with previous findings suggesting an ancient origin to the characteristic pattern of Native American mtDNA haplogroup distribution seen today among modern Caribbean islanders (Benn-Torres et al. 2015). However, I also identified several private or geographically restricted mtDNA haplotypes across the Caribbean. This indicates that drift and isolation led to some genetic differentiation between island populations. More broadly this finding suggests that female-mediated gene flow may have been limited and not essential to the maintenance of inter-island social networks.

Martínez-Cruzado (2010) predicted that nine Native American lineages currently found in modern Puerto Ricans had their proximate origin in pre-contact Puerto Rican populations. In the present sample, I identify two of these lineages, C1b2 and A2+16218, among the ancient inhabitants of Punta Candelero and Paso del Indio. Haplogroup C1b2 emerges as the characteristic motif of Ceramic Age Caribbean populations. This and

other findings support a South American origin for pre-contact Puerto Rican populations in agreement with previous genetics research (Bryc, Velez, et al. 2010, Gravel et al. 2013, Moreno-Estrada et al. 2013, Martínez-Cruzado 2002, 2010), and with extensive archaeological evidence (Siegel 2005, Rouse 1992, Chanlatte Baik 2003). However, the presence of several mtDNA lineages of unknown origin and the differentiation of the Caribbean mtDNA pool from that of extant South American groups suggests that genetic contributions from other regions such as ancient Mexico or Central America cannot be rejected. These findings provide tentative support for the multiple migration model for the peopling of the Antilles.

This research broadens understanding of the genetic relationship between modern Puerto Ricans and ancient island populations. Better characterization of the genetic diversity of the ancestral groups that have contributed genetic variation to modern Puerto Ricans may aid in development of genome association methods and guide future efforts at rare variant discovery among Puerto Rican biomedical cohorts. Given that a large amount of functional human genetic variation is population-specific, rare variants may be the primary genetic factors underlying complex phenotypes (Saint Pierre and Génin 2014). In addition, Native American "Taíno" ancestry is a central component of Puerto Rican national and ethnic identity (Oboler 1995, Ramos-Zayas 2003, Veran 2003). Cultural elements of pre-contact groups such as artwork and iconography are used today as expressions of contemporary Puerto Rican culture and symbols of identity and political empowerment (Duany 2001, De la Luz Rodriguez 2010, Oliver 2005). Self-identified Native American Taíno communities are found in Puerto Rico and the US Puerto Rican diaspora (Castanha 2010, Feliciano-Santos 2011, Duany 2001, Haslip-Viera 2014, Benn-

Torres 2014). Therefore the results of this work also have implications for the construction and experience of individual and group identity among contemporary Puerto Ricans (Laguer-Díaz 2013).

The conclusions of this research are limited by its almost exclusive focus on mtDNA, which although very informative, is still just a single locus that only provides information about female population history. Therefore, future research into this topic will focus on optimizing recovery of whole genome and Y-chromosome variants from the best-preserved ancient remains. Analysis of autosomal DNA will provide added resolution for detecting ancient admixture and inferring population relationships, and will also allow exploration of paternal relationships among ancient Caribbean groups. Additionally, future analyses will incorporate coalescent methods for direct reconstruction of demographic history, such as the extended Bayesian skyline plot, and modeling of several demographic scenarios for migration at the inter- and intra-island scale. At present, DNA extractions including sixty more dental remains from the three studied populations are underway.

# CHAPTER 5: CONCLUSION

This dissertation uses modern and ancient DNA to examine the role of migration and genetic admixture in the history and evolution of Caribbean populations. Specifically, the three main objectives of this research were to: (1) understand how differences in pre and post contact migration patterns and local selective pressures shaped human genomic variation across the Caribbean, (2) develop optimized methods for ancient DNA recovery with degraded remains from topical Caribbean environments, and (3) use paleogenomics data to reconstruct the population history of pre-contact indigenous groups from Puerto Rico and clarify their relationship to modern, admixed islanders.

In Chapter 2, I used the Illumina Infinium Multi-Ethnic Global Array to obtain autosomal genotypes from 55 self-identified Afro-Caribbeans from the Lesser Antilles islands of Grenada (n=6), St. Kitts (n=5), St. Lucia (n=15), Trinidad (n=19), and St. Vincent (n=10). I characterized patterns of genome-wide variation and ancestry in these groups and found they carry large proportions of African ancestry and smaller proportions of European, Native American, South and East Asian ancestry in their autosomal genomes. This pattern is significantly different from that observed across most of the Greater Antilles and also varies within the Lesser Antilles themselves. I also found strong signatures of sex-biased mating, which suggests that admixture may have ocurred largely between European males and females from other ethnicities. Additionally, I searched the autosomal genome for signatures of selection to asses whether recent adaptation to the environmental pressures of the Caribbean may have shaped these ancestry patterns. I found candidate targets of selection among olfactory genes potentially

associated with the major histocompatibility complex on chromosome 6, among other loci. But whether these signatures stem from selection before or after continental admixture still remains unclear. In summary, I found that post-contact demographic processes had a large impact on genetic diversity in the Lesser Antilles and that self-identified ethnic-cultural identity did not necessarily predict genetic ancestry.

In Chapter 3, I turn towards optimizing extraction methods for recovery of ancient and degraded DNA from tropical environments. I compare the performance of two protocols for ancient DNA extraction with skeletal remains. The first, Method D, was designed for recovery of short, degraded DNA fragments (Dabney et al. 2013) and the second, Method H, modifies the first by adding an initial wash, and an extended digestion and decalcification step (Warinner et al. 2014, Gamba et al. 2016). Both methods were tested on degraded human and chimpanzee remains excavated in the Caribbean and East Africa. Results indicate that both methods successfully recover ancient DNA and there is no significant difference in endogenous content or damage profiles in shotgun libraries built with either extract. But I found that Method D samples are enriched in smaller DNA fragments, have higher GC content and present a significantly higher enrichment factor after mitochondrial target capture. Since surviving DNA in ancient or historic remains from tropical contexts is likely to be extremely fragmented, these results suggest that at present, Method D is best suited for paleogenomics research with tropical ancient DNA.

Lastly, in Chapter 4, I used the insights from the previous chapter to guide selection of extraction methods. I sampled dental remains from 60 individuals (dated between A.D. 500–1300) collected from three pre-contact archaeological sites in Puerto Rico: Punta Candelero (n=19), Tibes (n=11) and Paso del Indio (n=30). Using in-solution

capture and next-generation sequencing, I obtained 27 complete mitochondrial genomes as well as partial autosomal genotypes from one individual. The mtDNA haplogroup distribution observed across all three communities, as well as the haplotypes represented, supported a mainly South American origin for pre-contact populations in Puerto Rico, but did not exclude possible contributions from other continental regions. In addition, I found two identical mtDNA haplotypes shared across pre and post contact Puerto Rican populations. This finding indicates at least some genetic continuity between the two groups regardless of the influence of post-contact demographic shifts. I also found limited instances of mtDNA lineage sharing between coeval pre-contact Caribbean populations, suggesting that female-mediated gene flow was not essential to Antillean social interaction. Unfortunately, the autosomal genotype data were inconclusive but additional efforts are currently ongoing to repeat whole genome enrichment on a subset of the Puerto Rican remains. The results of Chapter 3 agree with previous archaeological evidence and support the current prevailing view of the ancient Caribbean as a pluralistic culture area shaped by multiple migrations of continental Native American groups.

Overall, this work has led to a critical reassessment of existing models for the peopling of Puerto Rico and the Caribbean and a formal test of the hypotheses put forward by genetics research with modern Caribbean islanders (Keegan 2013, Martínez-Cruzado 2013). When integrated with historic, (bio)archaeological, and other lines of evidence, the findings reported here contribute to a more complete reconstruction of the population history of pre-contact Puerto Rico and of the extent and composition of inter-island gene flow during the Caribbean Ceramic Age. This in turn may help future archaeological interpretation regarding the role of kinship, residence patterns and

mobility in the maintenance of interaction networks across the pre-contact Antilles (Laffoon and Hoogland 2012, Mol 2013). Lastly, by sampling a large collection of radiocarbon dated skeletal remains (N=60), which span three archaeological sites and an extensive temporal period, this project is the largest study conducted with ancient DNA from pre-contact Caribbean populations to this day. The recovery of complete mitochondrial and partial autosomal data from the remains represents a significant gain in resolution over previous attempts, which were limited due to technological constraints, to examining small regions of the mtDNA genome (Mendisco et al. 2015, Lalueza-Fox et al. 2001, Lalueza-Fox et al. 2003).

A practical aspect of this dissertation was the optimization and systematic testing of protocols for increased aDNA recovery from skeletal remains of tropical environments. Ongoing research into this topic will contribute to pushing the current geographical boundaries of paleogenomics research and lessening the euro-centric bias of current ancient DNA studies (Hofreiter et al. 2015). With further optimization, the protocol designed during this research could also be extended for forensic applications, where surviving DNA may be as degraded as in ancient substrates due to extended postmortem intervals and corpse deposition in adverse preservation contexts (Imaizumi, Taniguchi, and Ogawa 2017, Soler et al. 2011, Milos et al. 2007).

Lastly, this dissertation has characterized how ancient indigenous groups, European colonial regimes, the African Slave Trade and modern labor movements have shaped the genomic diversity of extant Caribbean islanders. In addition to its anthropological and historical importance, such knowledge is also essential for informing the identification of funciontal, rare and medically relevant genetic variation in these

populations and for facilitating their inclusion in clinical genomics and individualized medicine efforts (Bustamante, De La Vega, and Burchard 2011, Popejoy and Fullerton 2016).

# REFERENCES

1000 Genomes Project, C., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. 2015. "A global reference for human genetic variation." *Nature* 526 (7571):68-74.

Achilli, A., U. A. Perego, C. M. Bravi, M. D. Coble, Q. P. Kong, S. R. Woodward, A. Salas, A. Torroni, and H. J. Bandelt. 2008. "The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies." *PLoS One* 3 (3):e1764.

Adler, C. J., W. Haak, D. Donlon, and A. Cooper. 2011. "Survival and recovery of DNA from ancient teeth and bones." *Journal of Archaeological Science* 38 (5):956-964.

Aird, D., M. G. Ross, W. S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke. 2011. "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries." *Genome Biology* 12 (2):R18.

Alexander, D. H., and K. Lange. 2011. "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation." *BMC Bioinformatics* 12:246.

Alexander, D. H., J. Novembre, and K. Lange. 2009. "Fast model-based estimation of ancestry in unrelated individuals." *Genome Research* 19 (9):1655-64.

Allentoft, M. E., M. Collins, D. Harker, J. Haile, C. L. Oskam, M. L. Hale, P. F. Campos, J. A. Samaniego, M. T. Gilbert, E. Willerslev, G. Zhang, R. P. Scofield, R. N. Holdaway, and M. Bunce. 2012. "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils." *Proc Biol Sci* 279 (1748):4724-33.

Alves-Silva, J., M. da Silva Santos, P. E. Guimaraes, A. C. Ferreira, H. J. Bandelt, S. D. Pena, and V. F. Prado. 2000. "The ancestry of Brazilian mtDNA lineages." *American Journal of Human Genetics* 67 (2):444-61.

Anderson, C. A., F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. 2010. "Data quality control in genetic case-control association studies." *Nature Protocols* 5 (9):1564-73.

Anderson-Córdova, K. 2005. "The Aftermath of Conquest: The Indians of Puerto Rico during the Early Sixteenth Century. ." In *Ancient Borinquen: Archaeology and ethnohistory of native Puerto Rico.*, edited by P. Siegel, 337-352. Tuscaloosa: University of Alabama Press.

Andrews, R. M., I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. 1999. "Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA." *Nature Genetics* 23 (2):147-147.

Andújar, C. 2006. *La presencia negra en Santo Domingo: Un enfoque ethnohistórico*. Santo Domingo: Editorial Letra Gráfica.

Auguie, B. gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.2.1.

Avila-Arcos, M. C., E. Cappellini, J. A. Romero-Navarro, N. Wales, J. V. Moreno-Mayar, M. Rasmussen, S. L. Fordyce, R. Montiel, J. P. Vielle-Calzada, E. Willerslev, and M. T. Gilbert. 2011. "Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA." *Scientific Reports* 1:74.

Ávila-Arcos, M. C., M. Sandoval-Velasco, H. Schroeder, M. L. Carpenter, A.-S. Malaspinas, N. Wales, F. Peñaloza, C. D. Bustamante, M. T. P. Gilbert, and M. Bunce. 2015. "Comparative performance of two whole-genome capture methodologies on ancient DNA Illumina libraries." *Methods in Ecology and Evolution* 6 (6):725-734.

Bahbahani, H., H. Clifford, D. Wragg, M. N. Mbole-Kariuki, C. Van Tassell, T. Sonstegard, M. Woolhouse, and O. Hanotte. 2015. "Signatures of positive selection in East African Shorthorn Zebu: A genome-wide single nucleotide polymorphism analysis." *Scientific Reports* 5:11729.

Bandelt, H.-J., P. Forster, and A. Rohl. 1999. "Median-Joining Networks for Inferring Intraspecific Phylogenies." *Molecular Biology and Evolution* 16 (1):37-48.

Bandelt, H.-J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. "Mitochondrial portraits of human populations using median networks." *Genetics* 141 (2):743-753.

Barbieri, C., P. Heggarty, L. Castri, D. Luiselli, and D. Pettener. 2011. "Mitochondrial DNA variability in the Titicaca basin: Matches and mismatches with linguistics and ethnohistory." *American Journal of Human Biology* 23 (1):89-99.

Bates, D., and M. Maechler. Matrix: Sparse and Dense Matrix Classes and Methods, R package version 1.2-7.1. https://cran.r-project.org/package=Matrix.

Behr, A. A., K. Z. Liu, G. Liu-Fang, P. Nakka, and S. Ramachandran. 2016. "pong: fast analysis and visualization of latent clusters in population genetic data." *Bioinformatics* 32 (18):2817-23.

Benn-Torres, J. 2014. "Prospecting the past: Genetic perspectives on the extinction and survival of indigenous poeples of the Caribbean." *New Genetics and Society* 33 (1):21-41.

Benn-Torres, J., C. Bonilla, C. M. Robbins, L. Waterman, T. Y. Moses, W. Hernandez, E. R. Santos, F. Bennett, W. Aiken, T. Tullock, K. Coard, A. Hennis, S. Wu, B. Nemesure, M. C. Leske, V. Freeman, J. Carpten, and R. A. Kittles. 2008. "Admixture and population stratification in African Caribbean populations." *Annals of Human Genetics* 72 (Pt 1):90-8.

Benn-Torres, J., M. B. Doura, S. O. Y. Keita, and R. Kittles. 2012. "Y Chromosome Lineages in Men of West African Descent." *PLoS One* 7 (1):e29687.

Benn-Torres, J., R. A. Kittles, and A. C. Stone. 2007. "Mitochondrial and Y chromosome diversity in the English-speaking Caribbean." *Annals of Human Genetics* 71 (Pt 6):782-90.

Benn-Torres, J., A. C. Stone, and R. Kittles. 2013. "An anthropological genetic perspective on Creolization in the Anglophone Caribbean." *American Journal of Physical Anthropology* 151 (1):135-43.

Benn-Torres, J., M. G. Vilar, G. A. Torres, J. B. Gaieski, R. Bharath Hernandez, Z. E. Browne, M. Stevenson, W. Walters, T. G. Schurr, and C. Genographic. 2015. "Genetic Diversity in the Lesser Antilles and Its Implications for the Settlement of the Caribbean Basin." *PLoS One* 10 (10):e0139192.

Berger, J., H. Kurahashi, Y. Takihara, K. Shimada, H. W. Brock, and F. Randazzo. 1999. "The human homolog of Sex comb on midleg (SCMH1) maps to chromosome 1p34." *Gene* 237 (1):185-191.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. 2004. "Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene." *The American Journal of Human Genetics* 74 (6):1111-1120.

Bhatia, G., A. Tandon, N. Patterson, M. C. Aldrich, C. B. Ambrosone, C. Amos, E. V. Bandera, S. I. Berndt, L. Bernstein, W. J. Blot, C. H. Bock, N. Caporaso, G. Casey, S. L. Deming, W. R. Diver, S. M. Gapstur, E. M. Gillanders, C. C. Harris, B. E. Henderson, S. A. Ingles, W. Isaacs, P. L. De Jager, E. M. John, R. A. Kittles, E. Larkin, L. H. McNeill, R. C. Millikan, A. Murphy, C. Neslund-Dudas, S. Nyante, M. F. Press, J. L. Rodriguez-Gil, B. A. Rybicki, A. G. Schwartz, L. B. Signorello, M. Spitz, S. S. Strom, M. A. Tucker, J. K. Wiencke, J. S. Witte, X. Wu, Y. Yamamura, K. A. Zanetti, W. Zheng, R. G. Ziegler, S. J. Chanock, C. A. Haiman, D. Reich, and A. L. Price. 2014. "Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture." *American Journal of Human Genetics* 95 (4):437-44.

Bien, S. A., G. L. Wojcik, N. Zubair, C. R. Gignoux, A. R. Martin, J. M. Kocarnik, L. W. Martin, S. Buyske, J. Haessler, R. W. Walker, I. Cheng, M. Graff, L. Xia, N. Franceschini, T. Matise, R. James, L. Hindorff, L. Le Marchand, K. E. North, C. A. Haiman, U. Peters, R. J. Loos, C. L. Kooperberg, C. D. Bustamante, E. E. Kenny, C. S. Carlson, and P. Study. 2016. "Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array." *PLoS One* 11 (12):e0167758.

Birth, K. 1997. "Most of us are family some of the time: Interracial unions and transracial kinship in Eastern Trinidad." *American Ethnologist* 24 (3):585-601.

Bodner, M., U. A. Perego, G. Huber, L. Fendt, A. W. Rock, B. Zimmermann, A. Olivieri, A. Gomez-Carballa, H. Lancioni, N. Angerhofer, M. C. Bobillo, D. Corach, S. R. Woodward, A. Salas, A. Achilli, A. Torroni, H. J. Bandelt, and W. Parson. 2012. "Rapid coastal spread of First Americans: novel insights from South America's Southern Cone mitochondrial genomes." *Genome Research* 22 (5):811-20.

Boessenkool, S., K. Hanghoj, H. M. Nistelberger, C. Der Sarkissian, A. T. Gondek, L. Orlando, J. H. Barrett, and B. Star. 2016. "Combining bleach and mild

predigestion improves ancient DNA recovery from bones." *Molecular Ecology Resources*.

Bolnick, D. A. 2011. "Continuity and Change in Anthropological Perspectives on Migration: Insights from Molecular Anthropology." In *Rethinking Anthropological Perspectives on Migration*, edited by G. S. Cabana and J. J. Clark, 263-277. Gainesville: University Press of Florida.

Bragulla, H. H., and D. G. Homberger. 2009. "Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia." *Journal of Anatomy* 214 (4):516-59.

Bramanti, B., M. G. Thomas, W. Haak, M. Unterlaender, P. Jores, K. Tambets, I. Antanaitis-Jacobs, M. N. Haidle, R. Jankauskas, C. J. Kind, F. Lueth, T. Terberger, J. Hiller, S. Matsumura, P. Forster, and J. Burger. 2009. "Genetic discontinuity between local hunter-gatherers and central Europe's first farmers." *Science* 326 (5949):137-40.

Braun, T., P. Voland, L. Kunz, C. Prinz, and M. Gratzl. 2007. "Enterochromaffin cells of the human gut: sensors for spices and odorants." *Gastroenterology* 132 (5):1890-901.

Briggs, A. W., U. Stenzel, Philip L. F. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, and S. Pääbo. 2007. "Patterns of damage in genomic DNA sequences from a Neandertal." *Proceedings of the National Academy of Sciences* 104 (37):14616–14621.

Brisbin, A., K. Bryc, J. Byrnes, F. Zakharia, L. Omberg, J. D. Degenhardt, A. Reynolds, H. Ostrer, J. G. Mezey, and C. Bustamante. 2012. "PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations." *Human Biology* 84 (4):343-364.

Brown, J. H. 2014. "Why are there so many species in the tropics?" *Journal of Biogeography* 41 (1):8-22.

Bryc, K., A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser, S. Williams, A. Froment, J.-M. Bodo, C. Wambebe, S. A. Tishkoff, and C. D. Bustamante. 2010. "Genome-wide patterns of population structure and admixture in West Africans

and African Americans." *Proceedings of the National Academy of Sciences* 107 (2):786-791.

Bryc, K., Eric Y. Durand, J. M. Macpherson, D. Reich, and Joanna L. Mountain. 2015. "The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States." *The American Journal of Human Genetics* 96 (1):37-53.

Bryc, K., C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C. D. Bustamante, and H. Ostrer. 2010. "Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations." *Proceedings of the National Academy of Sciences* 107 Suppl 2:8954-61.

Bryk, J., E. Hardouin, I. Pugach, D. Hughes, R. Strotmann, M. Stoneking, and S. Myles. 2008. "Positive Selection in East Asians for an EDAR Allele that Enhances NF-κB Activation." *PLoS One* 3 (5):e2209.

Bustamante, C. D., F. M. De La Vega, and E. G. Burchard. 2011. "Genomics for the world." *Nature* 475 (7355):163-165.

Buzas, M. A., L. S. Collins, and S. J. Culver. 2002. "Latitudinal difference in biodiversity caused by higher tropical rate of increase." *Proceedings of the National Academy of Sciences* 99 (12):7841–7843.

Cabana, G. 2011. "The problematic relationship between migration and culture change." In *Rethinking anthropological perspectives on migration*, edited by G. S. Cabana and J. J. Clark, 16-30. Gainesville: University Press of Florida.

Cadzow, M., J. Boocock, H. T. Nguyen, P. Wilcox, T. R. Merriman, and M. A. Black. 2014. "A bioinformatics workflow for detecting signatures of selection in genomic data." *Frontiers in Genetics* 5:293.

Callaghan, R. T. 2003. "Comments on the mainland origins of the Preceramic cultures of the Greater Antilles." *Latin American Antiquity* 14 (3):323-338.

Callaghan, R. T. 2013. "Archaeological Views of Caribbean Seafaring." In *The Oxford Handbook of Caribbean Archaeology*, edited by W. Keegan, C. L. Hofman and R. Rodríguez Ramos, 283-295. Oxford: Oxford University Press.

Cann, H. M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Huang, J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset, and L. L. Cavalli-Sforza. 2002. "A Human Genome Diversity Cell Line Panel." *Science* 296 (5566):261.

Cardena, M. M., A. Ribeiro-Dos-Santos, S. Santos, A. J. Mansur, A. C. Pereira, and C. Fridman. 2013. "Assessment of the relationship between self-declared ethnicity, mitochondrial haplogroups and genomic ancestry in Brazilian individuals." *PLoS One* 8 (4):e62005.

Cardona, A., L. Pagani, T. Antao, D. J. Lawson, C. A. Eichstaedt, B. Yngvadottir, M. T. Shwe, J. Wee, I. G. Romero, S. Raj, M. Metspalu, R. Villems, E. Willerslev, C. Tyler-Smith, B. A. Malyarchuk, M. V. Derenko, and T. Kivisild. 2014. "Genome-wide analysis of cold adaptation in indigenous Siberian populations." *PLoS One* 9 (5):e98076.

Carpenter, M. L., J. D. Buenrostro, C. Valdiosera, H. Schroeder, M. E. Allentoft, M. Sikora, M. Rasmussen, S. Gravel, S. Guillen, G. Nekhrizov, K. Leshtakov, D. Dimitrova, N. Theodossiev, D. Pettener, D. Luiselli, K. Sandoval, A. Moreno-Estrada, Y. Li, J. Wang, M. T. Gilbert, E. Willerslev, W. J. Greenleaf, and C. D. Bustamante. 2013. "Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries." *American Journal of Human Genetics* 93 (5):852-64.

Castanha, T. 2010. *The myth of indigenous Caribbean extinction: Continuity and reclamation in Boirkén (Puerto Rico)*. New York: Palgrave Macmillan.

Cavalli-Sforza, L. L., and M. W. Feldman. 2003. "The application of molecular genetic approaches to the study of human evolution." *Nature Genetics* 33 Suppl:266-75.

Chang, C. C., C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. "Second-generation PLINK: rising to the challenge of larger and richer datasets." *GigaScience* 4 (1):7.

Chanlatte Baik, L. A. 2003. "Agricultural societies in the Caribbean: The Greater Antilles, and the Bahamas." In *General History of the Caribbean, Vol. I. Autochtonous Societies*, 228-258. London: UNESCO Publishing.

Charlesworth, B. 2009. "Effective population size and patterns of molecular evolution and variation." *Nature Reviews Genetics* 10 (3):195-205.

Chimusa, E. R., A. Meintjies, M. Tchanga, N. Mulder, C. Seoighe, H. Soodyall, and R. Ramesar. 2015. "A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations." *PLoS Genetics* 11 (3):e1005052.

CIA. 2013. "The World Factbook 2013-14." Central Intelligence Agency. https://www.cia.gov/library/publications/the-world-factbook/index.html.

Clarke, C. 2011. "Demographic change and population movement." In *General History of the Caribbean. Volume IV: The Long Nineteenth Century: Nineteenth Century Transformations.*, edited by K. O. Laurence, 259-282. London: UNESCO Publishing.

Cobley, A. G. 1994. *Crossroads of Empire: The European-Caribbean Connection, 1492-1992*: Department of History, University of the West Indies.

Cook, N. D. 1998. *Born to Die: Disease and New World Conquest.* Cambridge: Cambridge University Press.

Cooper, A., and H. Poinar. 2000. "Ancient DNA: Do It Right or Not at All." *Science* 289 (5482):1139.

Coppa, A., A. Cucina, M. L. P. Hoogland, M. Lucci, F. Luna Calderón, R. G. A. M. Panhuysen, G. Tavarez Maria, R. Valcárcel Rojas, and R. Vargiu. 2008. "New evidence of two different migratory waves in the Circum-Caribbean area during the pre-Columbian period from the analysis of dental morphological traits." In *Crossing the borders: New methods and techniques in the study of archaeological materials from the Caribbean*, edited by C. L. Hofman, M. L. P. Hoogland and A. L. van Gijn, 195-213. Tuscaloosa: University of Alabama Press.

Crespo-Torres, E. 2000. "Estudio Comparativo Biocultural entre dos Poblaciones Prehistóricas en la Isla Puerto Rico: Punta Candelero y Paso del Indio."PhD

Dissertation, Instituto de Investigaciones Arqueológicas, Universidad Nacional Autónoma de México.

Cruxent, J. M., and I. Rouse. 1969. "Early man in the West Indies." *Scientific American* 221 (5):42-52.

Cruz-Davalos, D. I., B. Llamas, C. Gaunitz, A. Fages, C. Gamba, J. Soubrier, P. Librado, A. Seguin-Orlando, M. Pruvost, A. H. Alfarhan, S. A. Alquraishi, K. A. Al-Rasheid, A. Scheu, N. Beneke, A. Ludwig, A. Cooper, E. Willerslev, and L. Orlando. 2016. "Experimental conditions improving in-solution target enrichment for ancient DNA." *Molecular Ecology Resources*.

Cui, R., M. Schumer, and G. G. Rosenthal. 2016. "Admix'em: a flexible framework for forward-time simulations of hybrid populations with selection and mate choice." *Bioinformatics* 32 (7):1103-5.

Curet, L. A. 2004. "Island Archaeology and Units of Analysis in the Study of Caribbean Societies." In *Voyages of Discovery: The Archaeology of Islands*, edited by S. M. Fitzpatrick. Conneticut: Praeger Publishers.

Curet, L. A. 2005. *Caribbean Paleodemography: Population, culture history and sociopolitical processes in ancient Puerto Rico*. Tuscaloosa: University of Alabama Press.

Curet, L. A. 2014. "The Taino: Phenomena, Concepts, and Terms." *Ethnohistory* 61 (3):467-495.

Curet, L. A., and M. W. Hauser. 2011. "Introduction: Migration, Seafaring, and Cultural Contact in the Caribbean." In *Caribbean Archaeology and Ethnohistory : Islands at the Crossroads : Migration, Seafaring, and Interaction in the Caribbean*, edited by L. A. Curet and M. W. Hauser. Tuscaloosa: Unviersity of Alabama Press.

Curet, L. A., and L. M. Stringer. 2010. *Tibes: People, Power, and Ritual at the Center of the Cosmos*. Tuscaloosa: University of Alabama Press.

Curet, L. A., J. Torres, and M. Rodríguez. 2004. "Political and social history of Eastern Puerto Rico: the Ceramic age." In *Late Ceramic Age Societies in the Eastern*

*Caribbean*, edited by C. L. Hofman and A. Delpuech, 59-86. Oxford: Paris Monographs in American Archaeology 14.

Dabney, J., M. Knapp, I. Glocke, M. T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. Garcia, S. Paabo, J. L. Arsuaga, and M. Meyer. 2013. "Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments." *Proceedings of the National Academy of Sciences* 110 (39):15758-63.

Dabney, J., and M. Meyer. 2012. "Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries." *BioTechniques* 52 (2):87-94.

Dabney, J., M. Meyer, and S. Pääbo. 2013. "Ancient DNA damage." *Cold Spring Harbor Perspectives in Biology* 5 (7):a012567.

Daley, T., and A. D. Smith. 2013. "Predicting the molecular complexity of sequencing libraries." *Nature Methods* 10 (4):325-327.

Damgaard, P. B., A. Margaryan, H. Schroeder, L. Orlando, E. Willerslev, and M. E. Allentoft. 2015. "Improving access to endogenous DNA in ancient bones and teeth." *Scientific Reports* 5:11184.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. "The variant call format and VCFtools." *Bioinformatics* 27 (15):2156-2158.

Dannemann, M., A. M. Andres, and J. Kelso. 2016. "Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors." *American Journal of Human Genetics* 98 (1):22-33.

Daya, M., L. van der Merwe, C. R. Gignoux, P. D. van Helden, M. Möller, and E. G. Hoal. 2014. "Using multi-way admixture mapping to elucidate TB susceptibility in the South African Coloured population." *BMC Genomics* 15 (1):1021.

De la Luz Rodriguez, G. 2010. "Taino as a romantic term: Notes on the representation of the indigenous in Puerto Rican Archaeology and Ethnohistory." In *Bridging the*

*Divide : Indigenous Communities and Archaeology into the 21st Century*, edited by C. Phillips and H. Allen, 93-106. Walnut Creek: Left Coast Press.

de Leeuw, J., and P. Mair. 2009. "Multidimensional Scaling Using Majorization: SMACOF in R." *Journal of Statistical Software* 31 (3):1-30.

Deagle, B. E., J. P. Eveson, and S. N. Jarman. 2006. "Quantification of damage in DNA recovered from highly degraded samples--a case study on DNA in faeces." *Frontiers in Zoology* 3:11.

Delaneau, O., J. Marchini, and J.-F. Zagury. 2012. "A linear complexity phasing method for thousands of genomes." *Nat Meth* 9 (2):179-181.

Deng, L., A. Ruiz-Linares, S. Xu, and S. Wang. 2016. "Ancestry variation and footprints of natural selection along the genome in Latin American populations." *Scientific Reports* 6:21766.

Deschamps, M., G. Laval, M. Fagny, Y. Itan, L. Abel, J. L. Casanova, E. Patin, and L. Quintana-Murci. 2016. "Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes." *American Journal of Human Genetics* 98 (1):5-21.

Diptee, A. A. 2000. "Indian men, Afro-creole women: 'Casting' doubt on interracial sexual relationships in the late nineteenth-century Caribbean." *Immigrants & Minorities* 19 (3):1-24.

Duany, J. 2001. "Making Indians out of Blacks: The Revitalization of Taino identity in contemporary Puerto Rico." In *Taino Revival: Critical Perspectives on Puerto Rican Identity and Cultural Politics*, edited by G. Haslip-Viera, 55-82. Princeton: Markus Weiner Publishers.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. 2005. "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis." *Bioinformatics* 21 (16):3439-3440.

Durinck, S., P. T. Spellman, E. Birney, and W. Huber. 2009. "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." *Nat. Protocols* 4 (8):1184-1191.

Ehlers, A., S. Beck, S. Forbes, J. Trowsdale, A. Volz, R. M. Younger, and A. Ziegler. 2001. "MHC-linked olfactory receptor loci exhibit polymorphism and contribute to extended HLA/OR-Haplotypes." *Genome Research* 10:1968-1978.

Emery, L. S., K. M. Magnaye, A. W. Bigham, J. M. Akey, and M. J. Bamshad. 2015. "Estimates of continental ancestry vary widely among individuals with the same mtDNA haplogroup." *American Journal of Human Genetics* 96 (2):183-93.

Engerman, S. L., and B. W. Higman. 1997. "The demographic structure of the Caribbean slave societies in the eighteenth and nineteenth centuries." In *General History of the Caribbean. Volume III: The Slave Societies of the Caribbean.*, edited by F. W. Knight, 45-104. London: UNESCO Publishing.

Etkins, O. 2016. "Pick a box, any box: The influence of commercialized genetics on Caribbean biosociality." *Intersect* 9 (3):1-11.

Excoffier, L., and H. E. L. Lischer. 2010. "Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows." *Molecular Ecology Resources* 10:564-567.

Feliciano-Santos, S. 2011. "An Inconceivable Indigeneity: The Historical, Cultural, and Interactional Dimensions of Puerto Rican Taíno Activism. PhD Dissertation. University of Michigan." Doctor of Philosophy, Department of Anthropology.

Ferrer, I., P. Garcia-Esparcia, M. Carmona, E. Carro, E. Aronica, G. G. Kovacs, A. Grison, and S. Gustincich. 2016. "Olfactory Receptors in Non-Chemosensory Organs: The Nervous System in Health and Disease." *Frontiers in Aging Neuroscience* 8:163.

Figueiro, G., P. C. Hidalgo, and M. Sans. 2011. "Control region variability of haplogroup C1d and the tempo of the peopling of the Americas." *PLoS One* 6 (6):e20978.

Fontanez, R. 1991. "Restos Faunisticos y Explotacion del Medioambiente en Punta Candelero, Puerto Rico. Informe Preliminar." Proceedings of the 13th International Congress for Caribbean Archaeology, Curaçao.

Forster, P. 2004. "Ice Ages and the mitochondrial DNA chronology of human dispersals: a review." *Philosophical Transactions of the Royal Society B: Biological Sciences* 359 (1442):255-64; discussion 264.

Forte, M. C. 2006. *Indigenous resurgence in the contemporary Caribbean: Amerindian survival and revival*. New York: Peter Lang Publishing.

Fu, Q., A. Mittnik, P. L. Johnson, K. Bos, M. Lari, R. Bollongino, C. Sun, L. Giemsch, R. Schmitz, J. Burger, A. M. Ronchitelli, F. Martini, R. G. Cremonesi, J. Svoboda, P. Bauer, D. Caramelli, S. Castellano, D. Reich, S. Paabo, and J. Krause. 2013. "A revised timescale for human evolution based on ancient mitochondrial genomes." *Current Biology* 23 (7):553-9.

Gamba, C., K. Hanghoj, C. Gaunitz, A. H. Alfarhan, S. A. Alquraishi, K. A. Al-Rasheid, D. G. Bradley, and L. Orlando. 2016. "Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing." *Molecular Ecology Resources* 16 (2):459-69.

Gamba, C., E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes, V. Mattiangeli, L. Domboroczki, I. Kovari, I. Pap, A. Anders, A. Whittle, J. Dani, P. Raczky, T. F. Higham, M. Hofreiter, D. G. Bradley, and R. Pinhasi. 2014. "Genome flux and stasis in a five millennium transect of European prehistory." *Nature Communications* 5:5257.

Gautier, M., and M. Naves. 2011. "Footprints of selection in the ancestral admixture of a New World Creole cattle breed." *Molecular Ecology* 20 (15):3128-43.

Ghasemi, A., and S. Zahediasl. 2012. "Normality tests for statistical analysis: a guide for non-statisticians." *Internaiontal Journal of Endocrinology and Metabolism* 10 (2):486-9.

Gilad, Y., C. D. Bustamante, D. Lancet, and S. Paabo. 2003. "Natural selection on the olfactory receptor gene family in humans and chimpanzees." *American Journal of Human Genetics* 73 (3):489-501.

Gilad, Y., and D. Lancet. 2003. "Population Differences in the Human Functional Olfactory Repertoire." *Molecular Biology and Evolution* 20 (3):307-314.

Gilbert, M. T. P., A. J. Hansen, E. Willerslev, G. Turner-Walker, and M. Collins. 2006. "Insights into the processes behind the contamination of degraded human teeth and bone samples with exogenous sources of DNA." *International Journal of Osteoarchaeology* 16 (2):156-164.

Giles, R. E., H. Blanc, H. M. Cann, and D. C. Wallace. 1980. "Maternal inheritance of human mitochondrial DNA." *Proceedings of the National Academy of Sciences* 77 (11):6715-6719.

Giner, G., and G. Smyth. 2016. "statmod: probability calculations for the inverse Gaussian distribution." *R Journal* 8 (1):339-351.

Ginolhac, A., M. Rasmussen, M. T. P. Gilbert, E. Willerslev, and L. Orlando. 2011. "mapDamage: testing for damage patterns in ancient DNA sequences." *Bioinformatics* 27 (15):2153-2155.

Ginolhac, A., J. Vilstrup, J. Stenderup, M. Rasmussen, M. Stiller, B. Shapiro, G. Zazula, D. Froese, K. E. Steinmann, J. F. Thompson, K. A. Al-Rasheid, T. M. Gilbert, E. Willerslev, and L. Orlando. 2012. "Improving the performance of true single molecule sequencing for ancient DNA." *BMC Genomics* 13:177.

Glusman, G., I. Yanai, I. Rubin, and D. Lancet. 2001. "The complete human olfactory subgenome." *Genome Research* 11 (5):685-702.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum. 2009. "Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing." *Nature Biotechnology* 27 (2):182-189.

Gomez-Carballa, A., L. Catelli, J. Pardo-Seco, F. Martinon-Torres, L. Roewer, C. Vullo, and A. Salas. 2015. "The complete mitogenome of a 500-year-old Inca child mummy." *Scientific Reports* 5:16462.

Gonzalez-Martin, A., A. Gorostiza, L. Regalado-Liu, S. Arroyo-Pena, S. Tirado, I. Nuno-Arana, R. Rubi-Castellanos, K. Sandoval, M. D. Coble, and H. Rangel-Villalobos. 2015. "Demographic History of Indigenous Populations in Mesoamerica Based on mtDNA Sequence Data." *PLoS One* 10 (8):e0131791.

Gorostiza, A., V. Acunha-Alonzo, L. Regalado-Liu, S. Tirado, J. Granados, D. Sámano, H. Rangel-Villalobos, and A. González-Martín. 2012. "Reconstructing the History of Mesoamerican Populations through the Study of the Mitochondrial DNA Control Region." *PLoS One* 7 (9):e44666.

Gourraud, P.-A., P. Khankhanian, N. Cereb, S. Y. Yang, M. Feolo, M. Maiers, J. D. Rioux, S. Hauser, and J. Oksenberg. 2014. "HLA Diversity in the 1000 Genomes Dataset." *PLoS One* 9 (7):e97282.

Gravel, S., F. Zakharia, A. Moreno-Estrada, J. K. Byrnes, M. Muzzio, J. L. Rodriguez-Flores, E. E. Kenny, C. R. Gignoux, B. K. Maples, W. Guiblet, J. Dutil, M. Via, K. Sandoval, G. Bedoya, P. Genomes, T. K. Oleksyk, A. Ruiz-Linares, E. G. Burchard, J. C. Martínez-Cruzado, and C. D. Bustamante. 2013. "Reconstructing Native American migrations from whole-genome and whole-exome data." *PLoS Genetics* 9 (12):e1004023.

Guardado-Estrada, M., E. Juarez-Torres, I. Medina-Martinez, A. Wegier, A. Macias, G. Gomez, F. Cruz-Talonia, E. Roman-Bassaure, D. Pinero, S. Kofman-Alfaro, and J. Berumen. 2009. "A great diversity of Amerindian mitochondrial DNA ancestry is present in the Mexican mestizo population." *Journal of Human Genetics* 54 (12):695-705.

Günther, T., C. Valdiosera, H. Malmstrom, I. Ureña, R. Rodriguez-Varela, Ó. O. Sverrisdóttir, E. A. Daskalaki, P. Skoglund, T. Naidoo, E. M. Svensson, J. M. Bermudez de Castro, E. Carbonell, M. Dunn, J. Storå, E. Iriarte, J. L. Arsuaga, J.-M. Carretero, A. Götherström, and M. Jakobsson. 2015. "Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques." *Proceedings of the National Academy of Sciences* 112 (38):1917–11922.

Gutierrez-Garcia, T. A., E. Vazquez-Dominguez, J. Arroyo-Cabrales, M. Kuch, J. Enk, C. King, and H. N. Poinar. 2014. "Ancient DNA and the tropics: a rodent's tale." *Biology Letters* 10 (6).

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Banffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. A. Rojo Guerra, C. Roth, A. Szecsenyi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, and D. Reich. 2015. "Massive migration from the steppe was a source for Indo-European languages in Europe." *Nature* 522 (7555):207-211.

Hansen, H. B., P. B. Damgaard, A. Margaryan, J. Stenderup, N. Lynnerup, E. Willerslev, and M. E. Allentoft. 2017. "Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum." *PLoS One* 12 (1):e0170940.

Harris, P. 1973. "Preliminary Report on Banwari-Trace: A Preceramic site in Trinidad." Proceedings of the 4th International Congress for the Study of Pre-Columbian Cultures of the Lesser Antilles. , St. Lucia.

Haslip-Viera, G. 2001. *Taino Revival: Critical prespectives on Puerto Rican identity and cultural politics*. Princetom: Markus Weiner Publishers.

Haslip-Viera, G. 2006. "The politics of Taíno revivalism: The insignificance of Amerindian mtDNA in the population history of Puerto Ricans. A comment on recent research." *Centro Journal* 18:260-275.

Haslip-Viera, G. 2014. *Race, Identity and Indigenous Politics: Puerto Rican Neo-Tainos in the Diaspora and the Island*. New York: Latino Studies Press.

Head, S. R., H. K. Komori, S. A. LaMere, T. Whisenant, F. Van Nieuwerburgh, D. R. Salomon, and P. Ordoukhanian. 2014. "Library construction for next-generation sequencing: overviews and challenges." *BioTechniques* 56 (2):61-4, 66, 68, passim.

Hedrick, P. W. 2011. *Genetics of populations*. Fourth ed. Massachusetts: Jones and Bartlett Publishers.

Heintzman, P. D., G. D. Zazula, J. A. Cahill, A. V. Reyes, R. D. MacPhee, and B. Shapiro. 2015. "Genomic Data from Extinct North American Camelops Revise Camel Evolutionary History." *Molecular Biology and Evolution* 32 (9):2433-40.

Helgason, A., C. Lalueza-Fox, S. Ghosh, S. Sampietro, E. Gigli, A. Baker, J. Bertranpetit, L. Arnadóttir, U. Þorsteinsdottir, and K. Stefánsson. 2009. "Sequences From First Settlers Reveal Rapid Evolution in Icelandic mtDNA Pool." *PLoS Genetics* 5 (1):e1000343.

Hellenthal, G., G. B. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. 2014. "A genetic atlas of human admixture history." *Science* 343 (6172):747-51.

Hennig, C. fpc: Flexible Procedures for Clustering, R package version 2.1-10. https://cran.r-project.org/package=fpc.

Hervé, M. RVAideMemoire: Diverse Basic Statistical and Graphical Functions, R package version 0.9-62. https://cran.r-project.org/package=RVAideMemoire.

Heuman, G. 1997. "The social structure of the slave societies in the Caribbean." In *General History of the Caribbean. Volume III: The Slave Societies of the Caribbean.*, edited by F. W. Knight, 138-168. London: UNESCO Publishing.

Ho, S. Y., and M. T. Gilbert. 2010. "Ancient mitogenomics." *Mitochondrion* 10 (1):1-11.

Hoetnik, H. 1985. ""Race" and color in the Caribbean." In *Caribbean Contours*, edited by S. W. Mintz and S. Price, 55-85. Baltimore: Johns Hopkins University Press.

Hofman, C. L., A. Boomert, A. J. Bright, M. L. P. Hoogland, S. Knippenberg, and A. V. M. Samson. 2011. "Ties with the Homelands: Archipelagic Interaction and the Enduring Role ofthe South and Central American Mainlands in the Pre-Columbian Lesser Antilles." In *Caribbean Archaeology and Ethnohistory : Islands at the Crossroads : Migration, Seafaring, and Interaction in the Caribbean.* , edited by L. A. Curet and M. W. Hauser, 73-86. Tuscaloosa: University of Alabama Press.

Hofreiter, M., J. L. Paijmans, H. Goodchild, C. F. Speller, A. Barlow, G. G. Fortes, J. A. Thomas, A. Ludwig, and M. J. Collins. 2015. "The future of ancient DNA: Technical advances and conceptual shifts." *Bioessays* 37 (3):284-93.

Hofreiter, M., D. Serre, H. Poinar, M. Kuch, and S. Pääbo. 2001. "Ancient DNA." *Nature Reviews Genetics* 2.

Homburger, J. R., A. Moreno-Estrada, C. R. Gignoux, D. Nelson, E. Sanchez, P. Ortiz-Tello, B. A. Pons-Estel, E. Acevedo-Vasquez, P. Miranda, C. D. Langefeld, S. Gravel, M. E. Alarcon-Riquelme, and C. D. Bustamante. 2015. "Genomic Insights into the Ancestry and Demographic History of South America." *PLoS Genetics* 11 (12):e1005602.

Hoogland, M. L. P., C. L. Hofman, and R. G. A. M. Panhuysen. 2010. "Interisland Dynamics: Evidence for Human Mobility at the site of Anse a la Gourde, Guadeloupe." In *Island shores, distant pasts: Archaeological and biological approaches to the Pre-Columbian settlement of the Caribbean*, edited by S. Fitzpatrick and A. Ross, 148-162. Gainesville: University Press of Florida.

Höss, M., and S. Pääbo. 1993. "DNA extraction from Pleistocene bones by a silica-based purification method." *Nucleic Acids Research* 21 (16):3913-3914.

Huerta-Sanchez, E., X. Jin, Asan, Z. Bianba, B. M. Peter, N. Vinckenbosch, Y. Liang, X. Yi, M. He, M. Somel, P. Ni, B. Wang, X. Ou, Huasang, J. Luosang, Z. X. Cuo, K. Li, G. Gao, Y. Yin, W. Wang, X. Zhang, X. Xu, H. Yang, Y. Li, J. Wang, J. Wang, and R. Nielsen. 2014. "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA." *Nature* 512 (7513):194-7.

Hunley, K., K. Gwin, and B. Liberman. 2016. "A Reassessment of the Impact of European Contact on the Structure of Native American Genetic Diversity." *PLoS One* 11 (8):e0161018.

Hutt, S., C. M. Blanco, and O. Varmer. 1999. *Heritage Resources Law: Protecting the Archeological and Cultural Environment*. USA: John Wiley & Sons.

Illumina, I. 2015. Infinium Expanded Multi-Ethnic Genotyping Array (MEGA-EX) Data Sheet.

Imaizumi, K., K. Taniguchi, and Y. Ogawa. 2017. "DNA survival and physical and histological properties of heat-induced alterations in burnt bones." *International Journal of Legal Medicine* 128 (3):439-446.

Jansson, E., J. Harmoinen, M. Ruokonen, and J. Aspi. 2014. "Living on the edge: reconstructing the genetic history of the Finnish wolf population." *BMC Evolutionary Biology* 14 (64).

Jeong, C., G. Alkorta-Aranburu, B. Basnyat, M. Neupane, D. B. Witonsky, J. K. Pritchard, C. M. Beall, and A. Di Rienzo. 2014. "Admixture facilitates genetic adaptations to high altitude in Tibet." *Nature Communications* 5:3281.

Jin, W., S. Xu, H. Wang, Y. Yu, Y. Shen, B. Wu, and L. Jin. 2012. "Genome-wide detection of natural selection in African Americans pre- and post-admixture." *Genome Research* 22 (3):519-27.

Jónsson, H., A. Gonolhac, M. Schubert, P. Johnson, and L. Orlando. 2013. "mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters." *Bioinformatics* 29 (13):1682-1684.

Kaestle, F. A., and K. A. Horsburgh. 2002. "Ancient DNA in anthropology: Methods, applications, and ethics." *American Journal of Physical Anthropology* 119 (S35):92-130.

Katoh, K., and D. M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4):772-780.

Keegan, W. 1995. "Modeling Dispersal in the Prehistoric West Indies." *World Archaeology* 26 (3):400-420.

Keegan, W. F. 2006. "Archaic Influences in the Origins and Development of Taino Societies." *Caribbean Journal of Science,* 42 (1):1-10.

Keegan, W. F. 2013. "Caribbean Islands: Archaeology." In *The Global Prehistory of Human Migration*, edited by I. Ness and P. Bellwood, 376-383. United Kingdom: Wiley-Blackwell.

Keegan, W. F., and M. D. Maclachlan. 1989. "The Evolution of Avunculocal Chiefdoms: A reconstruction of Taino Kinship and politics." *American Anthropologist* 91:613-630.

Keegan, W. F., and R. Rodriguez Ramos. 2005. "Archaic origins of the Classic Tainos." Proceedings of the 21st International Congress for Caribbean Archaeology, Trinidad.

Kehlmaier, C., A. Barlow, A. K. Hastings, M. Vamberger, J. L. Paijmans, D. W. Steadman, N. A. Albury, R. Franz, M. Hofreiter, and U. Fritz. 2017. "Tropical ancient DNA reveals relationships of the extinct Bahamian giant tortoise Chelonoidis alburyorum." *Proc Biol Sci* 284 (1846).

Khan, I., E. Maldonado, V. Vasconcelos, S. J. O'Brien, W. E. Johnson, and A. Antunes. 2014. "Mammalian keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and adaptation to terrestrial and aquatic environments." *BMC Genomics* 15 (1):779.

Kidd, J. M., S. Gravel, J. Byrnes, A. Moreno-Estrada, S. Musharoff, K. Bryc, J. D. Degenhardt, A. Brisbin, V. Sheth, R. Chen, S. F. McLaughlin, H. E. Peckham, L. Omberg, C. A. Bormann Chung, S. Stanley, K. Pearlstein, E. Levandowsky, S.

Acevedo-Acevedo, A. Auton, A. Keinan, V. Acuna-Alonzo, R. Barquera-Lozano, S. Canizales-Quinteros, C. Eng, E. G. Burchard, A. Russell, A. Reynolds, A. G. Clark, M. G. Reese, S. E. Lincoln, A. J. Butte, F. M. De La Vega, and C. D. Bustamante. 2012. "Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation." *American Journal of Human Genetics* 91 (4):660-71.

Kilman, R. M. 2016. Encyclopedia of Evolutionary Biology. USA: Academic Press.

Kircher, M., S. Sawyer, and M. Meyer. 2012. "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform." *Nucleic Acids Research* 40 (1):e3.

Knight, F. W. 1997a. "Introduction." In *General History of the Caribbean. Volume III: The Slave Societies of the Caribbean.*, edited by F. W. Knight, 1-8. London: UNESCO Publishing.

Knight, F. W. 1997b. "Pluralism, creolization and culture." In *General History of the Caribbean. Volume III: The slave societies of the Caribbean*, edited by F. W. Knight, 271-286. London: UNESCO Publishing.

Korlevic, P., T. Gerber, M. T. Gansauge, M. Hajdinjak, S. Nagel, A. Aximu-Petri, and M. Meyer. 2015. "Reducing microbial and human contamination in DNA extractions from ancient bones and teeth." *BioTechniques* 59 (2):87-93.

Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel. 2006. "World Map of the Köppen-Geiger climate classification updated." *Meteorologische Zeitschrift* 15 (3):259-263.

Krause, J., A. W. Briggs, M. Kircher, T. Maricic, N. Zwyns, A. Derevianko, and S. Pääbo. 2010. "A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia." *Current Biology* 20 (3):231-236.

Kruskal, J. B., and M. Wish. 1978. *Multidimensional Scaling*, *Series Quantitative Applications in the Social Sciences, Number 11.* : Sage Publications.

Kumar, S., C. Bellis, M. Zlojutro, P. E. Melton, J. Blangero, and J. E. Curran. 2011. "Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins." *BMC Evolutionary Biology* 11 (293).

Laffoon, J. E. 2013. "Paleomobility Research in Caribbean Contexts: New Perspectives from Isotope Analysis." In *The Oxford Handbook of Caribbean Archaeoloy*, edited by W. F. Keegan, C. L. Hofman and R. Rodriguez Ramos, 418-435. Oxford: Oxford University Press.

Laffoon, J. E., and M. L. P. Hoogland. 2012. "Migration and mobility in the circum-Caribbean: Integrating archaeology and isotopic analysis." In *Topoi. Berlin Studies of the Ancient World/ Topoi. Berliner Studien der Alten Welt : Population Dynamics in Prehistory and Early History : New Approaches Using Stable Isotopes and Genetics*, edited by J. Burger, E. Kaiser and W. Schier. Munchen, DEU: Walter de Gruyter Burger, Joachim, Kaiser, Elke, Schier, Wolfram.

Laffoon, J. E., E. Plomp, G. R. Davies, M. L. P. Hoogland, and C. L. Hofman. 2013. "The Movement and Exchange of Dogs in the Prehistoric Caribbean: An Isotopic Investigation." *International Journal of Osteoarchaeology* 25 (4):454-465.

Laffoon, J. E., R. Rodriguez Ramos, L. Chanlatte Baik, Y. Narganes Storde, M. Rodríguez Lopez, G. R. Davies, and C. L. Hofman. 2014. "Long-distance exchange in the precolonial Circum-Caribbean: A multi-isotope study of animal tooth pendants from Puerto Rico." *Journal of Anthropological Archaeology* 35:220-233.

Laguer-Díaz, C. 2013. "The Construction of an Identity and the Politics of Remembering." In *The Oxford Handbook of Caribbean Archaeology*, edited by W. F. Keegan, C. L. Hofman and R. Rodriguez Ramos, 557-567. Oxford: Oxford University Press.

Lalueza-Fox, C., F. L. Calderón, F. Calafell, B. Morera, and J. Bertranpetit. 2001. "MtDNA from extinct Tainos and the peopling of the Caribbean." *Annals of Human Genetics* 65 (Pt 2):137-51.

Lalueza-Fox, C., M. T. P. Gilbert, A. J. Martinez-Fuentes, F. Calafell, and J. Bertranpetit. 2003. "Mitochondrial DNA from pre-Columbian Ciboneys from Cuba and the prehistoric colonization of the Caribbean." *American Journal of Physical Anthropology* 121 (2):97-108.

Laurence, K. O. 2011. "The importation of labour and the contract systems." In *General History of the Caribbean. Volume IV: The Long Nineteenth Century: Nineteenth Century Transformations.*, edited by K. O. Laurence, 191-222. London: UNESCO Publishing.

Lee, E. J., and D. A. Merriwether. 2015. "Identification of Whole Mitochondrial Genomes from Venezuela and Implications on Regional Phylogenies in South America." *Human Biology* 87 (1):29-38.

Lee-Loy, A.-M. 2004. "Kissing the cross: Nineteenth-century representations of Chinese and Indian immigrants in British Guiana and Trinidad." In *The Chinese in the Caribbean*, edited by A. Wilson, 25-39. Princeton: Markus Wiener Publishers.

Levine, S. S., A. Weiss, H. Erdjument-Bromage, Z. Shao, P. Tempst, and R. E. Kingston. 2002. "The Core of the Polycomb Repressive Complex Is Compositionally and Functionally Conserved in Flies and Humans." *Molecular and Cellular Biology* 22 (17):6070-6078.

Lewis, C. M., Jr., B. Lizarraga, R. Y. Tito, P. W. Lopez, G. C. Iannacone, A. Medina, R. Martinez, S. I. Polo, A. F. De La Cruz, A. M. Caceres, and A. C. Stone. 2007. "Mitochondrial DNA and the peopling of South America." *Human Biology* 79 (2):159-78.

Li, H., and R. Durbin. 2009. "Fast and accurate short read alignment with Burrows–Wheeler transform." *Bioinformatics* 25 (14):1754-1760.

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

Librado, P., and J. Rozas. 2009. "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data." *Bioinformatics* 25 (11):1451-1452.

Lindahl, T. 1993. "Instability and decay of the primary structure of DNA." *Nature* 362:709-715.

Lischer, H. E. L., and L. Excoffier. 2012. "PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs." *Bioinformatics* 28 (2):298-299.

Liu, Y., T. Nyunoya, S. Leng, S. A. Belinsky, Y. Tesfaigzi, and S. Bruse. 2013. "Softwares and methods for estimating genetic ancestry in human populations." *Human Genomics* 7 (1):1.

Llamas, B., L. Fehren-Schmitz, G. Valverde, J. Soubrier, S. Mallick, N. Rohland, S. Nordenfelt, C. Valdiosera, S. M. Richards, A. Rohrlach, M. I. B. Romero, I. F. Espinoza, E. T. Cagigao, L. W. Jiménez, K. Makowski, I. S. L. Reyna, J. M. Lory, J. A. B. Torrez, M. A. Rivera, R. L. Burger, M. C. Ceruti, J. Reinhard, R. S. Wells, G. Politis, C. M. Santoro, V. G. Standen, C. Smith, D. Reich, S. Y. W. Ho, A. Cooper, and W. Haak. 2016. "Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas." *Science Advances* 2 (4).

Look Lai, W. 2004. "The Chinese indenture system in the British West Indies and its aftermath." In *The Chinese in the Caribbean*, edited by A. Wilson, 3-24. Princeton: Markus Wiener Publishers.

Lopez, K. 2004. ""One brings another": The formation of early twentieth century Chinese migrant communities in Cuba." In *The Chinese in the Caribbean*, edited by A. Wilson, 93-127. Princeton: Markus Wiener Publishers.

Ma, J., and C. I. Amos. 2012. "Principal components analysis of population admixture." *PLoS One* 7 (7):e40115.

Malnic, B., P. A. Godfrey, and L. B. Buck. 2004. "The human olfactory receptor gene family." *Proceedings of the National Academy of Sciences* 101 (8):2584-2589.

Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante. 2013. "RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference." *American Journal of Human Genetics* 93 (2):278-88.

Marcheco-Teruel, B., E. J. Parra, E. Fuentes-Smith, A. Salas, H. N. Buttenschon, D. Demontis, M. Torres-Espanol, L. C. Marin-Padron, E. J. Gomez-Cabezas, V. Alvarez-Iglesias, A. Mosquera-Miguel, A. Martinez-Fuentes, A. Carracedo, A. D. Borglum, and O. Mors. 2014. "Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers." *PLoS Genetics* 10 (7):e1004488.

Mardis, E. R. 2008. "Next-generation DNA sequencing methods." *Annual Review of Genomics and Human Genetics* 9:387-402.

Mardulyn, P. 2012. "Trees and ⁄or networks to display intraspecific DNA sequence variation?" *Molecular Ecology* 21 (3385–3390).

Maricic, T., M. Whitten, and S. Paabo. 2010. "Multiplexed DNA sequence capture of mitochondrial genomes using PCR products." *PLoS One* 5 (11):e14004.

Martin-Fragachan, G. 1999. "Intellectual, artistic and ideological aspects of cultures in the New World." In *General History of the Caribbean. Volume II: New Societies: The Caribbean in the Long Sixteenth Century*, edited by P. C. Emmer, 247-307. London: UNESCO Publishing.

Martínez-Cruzado, J. C. 2002. "The Use of Mitochondrial DNA to Discover Pre-Columbian Migrations to the Caribbean: Results for Puerto Rico and Expectations for the Dominican Republic." *KACIKE: Journal of Caribbean Amerindian History and Anthropology*:1-11.

Martínez-Cruzado, J. C. 2010. "The History of Amerindian mitochondrial DNA lineages in Puerto Rico." In *Island shores, distant pasts: Archaeological and biological approaches to the Pre-Columbian settlement of the Caribbean*, edited by S. Fitzpatrick and A. Ross, 54-80. Gainesville: University Press of Florida.

Martínez-Cruzado, J. C. 2013. "The DNA evidence for the human colonization and spread across the Americas: Implications for the peopling of the Caribbean." In *The Oxford Handbook of Caribbean Archaeology*, edited by W. F. Keegan, C. L. Hofman and R. Rodríguez Ramos, 470-485. Oxford: Oxford University Press.

Martinez-Cruzado, J. C., G. Toro-Labrador, V. Ho-Fung, M. Estevez-Montero, A. Lobaina-Manzanet, D. A. Padovani-Claudio, H. Sanchez-Cruz, P. Ortiz-Bermudez, and A. Sanchez-Crespo. 2001. "Mitochondrial DNA Analysis Reveals Substantial Native American Ancestry in Puerto Rico." *Human Biology* 73 (4):491-511.

Martínez-Cruzado, J. C., G. Toro-Labrador, J. Viera-Vera, M. Y. Rivera-Vega, J. Startek, M. Latorre-Esteves, A. Roman-Colon, R. Rivera-Torres, I. Y. Navarro-Millan, E. Gomez-Sanchez, H. Y. Caro-Gonzalez, and P. Valencia-Rivera. 2005. "Reconstructing the population history of Puerto Rico by means of mtDNA phylogeographic analysis." *American Journal of Physical Anthropology* 128 (1):131-55.

Mata-Míguez, J., L. Overholtzer, E. Rodríguez-Alegría, B. M. Kemp, and D. A. Bolnick. 2012. "The genetic impact of aztec imperialism: Ancient mitochondrial DNA evidence from Xaltocan, Mexico." *American Journal of Physical Anthropology* 149 (4):504-516.

McRae, J. F., J. D. Mainland, S. R. Jaeger, K. A. Adipietro, H. Matsunami, and R. D. Newcomb. 2012. "Genetic Variation in the Odorant Receptor OR2J3 Is Associated with the Ability to Detect the "Grassy" Smelling Odor, cis-3-hexen-1-ol." *Chemical Senses* 37 (7):585-593.

Mendisco, F., M. H. Pemonge, E. Leblay, T. Romon, G. Richard, P. Courtaud, and M. F. Deguilloux. 2015. "Where are the Caribs? Ancient DNA from ceramic period human remains in the Lesser Antilles." *Philosophical Transactions of the Royal Society B: Biological Sciences* 370 (1660):20130388.

Mendizabal, I., K. Sandoval, G. Berniell-Lee, F. Calafell, A. Salas, A. Martinez-Fuentes, and D. Comas. 2008. "Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba." *BMC Evolutionary Biology* 8:213.

Metzker, M. L. 2010. "Sequencing technologies - the next generation." *Nature Reviews Genetics* 11 (1):31-46.

Meyer, M., J. L. Arsuaga, C. de Filippo, S. Nagel, A. Aximu-Petri, B. Nickel, I. Martinez, A. Gracia, J. M. Bermudez de Castro, E. Carbonell, B. Viola, J. Kelso, K. Prufer, and S. Paabo. 2016. "Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins." *Nature* 531 (7595):504-7.

Meyer, M., Q. Fu, A. Aximu-Petri, I. Glocke, B. Nickel, J. L. Arsuaga, I. Martinez, A. Gracia, J. M. de Castro, E. Carbonell, and S. Paabo. 2014. "A mitochondrial genome sequence of a hominin from Sima de los Huesos." *Nature* 505 (7483):403-6.

Meyer, M., and M. Kircher. 2010. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing." *Cold Spring Harbor Protocols* 6:1-10.

Meyer, S., G. Weiss, and A. Von Haeseler. 1999. "Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA." *Genetics* 152:1103-1110.

Mi, H., X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P. D. Thomas. 2016. "PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements." *Nucleic Acids Research*.

Mi, H., A. Murunganujan, J. T. Casagrande, and P. D. Thomas. 2013. "Large-scale gene function analysis with the PANTHER classification system." *Nature Protocols* 8 (8):1551-1566.

Miljkovic-Gacic, I., R. E. Ferrell, A. L. Patrick, C. M. Kammerer, and C. H. Bunker. 2005. "Estimates of African, European and Native American ancestry in Afro-Caribbean men on the island of Tobago." *Human Heredity* 60 (3):129-33.

Milne, I., G. Stephen, M. Bayer, P. J. A. Cock, L. Pritchard, L. Cardle, P. D. Shaw, and D. Marshall. 2013. "Using Tablet for visual exploration of second-generation sequencing data." *Briefings in Bioinformatics* 14 (2):193-202.

Milos, A., A. Selmanovic, L. Smajilovic, R. Huel, C. Katzmarzyck, A. Rizvic, and T. Parsons. 2007. "Success rates of nuclear short tandem repeat typing from different skeletal elements." *Croatian Medical Journal* 48 (4):486-493.

Mizuno, F., J. Gojobori, L. Wang, and K. Onis. 2014. "Complete mitogenome analysis of indigenous populations in Mexico: its relevance for the origin of Mesoamericans." *Journal of Human Genetics* 59 (7):359-367.

Mohammed, P. 2002. *Gender negotiations among Indians in Trinidad 1917-1947*. New York: Palgrave Institute of Social Studies.

Mol, A. A. 2013. "Studying Pre-Columbian Interaction Networks: Mobility and Exchange." In *The Oxford Handbook of Caribbean Archaeology*, edited by W. F. Keegan, C. L. Hofman and R. Rodriguez Ramos, 329-346. Oxford: Oxford University Press.

Moll, R., M. Divo, and L. Langbein. 2008. "The human keratins: biology and pathology." *Histochemistry and Cell Biology* 129 (6):705-33.

Montinaro, F., G. B. Busby, V. L. Pascali, S. Myers, G. Hellenthal, and C. Capelli. 2015. "Unravelling the hidden ancestry of American admixed populations." *Nature Communications* 6:6596.

Moore, D. S., G. P. McCabe, and B. A. Craig. 2009. *Introduction to the Practice of Statistics*. Sixth ed. New York: W.H. Freedman.

Moreno Fraginals, M. 1977. *Africa en América Latina*. Mexico: Siglo XXI Editores.

Moreno-Estrada, A., F. Casals, A. Ramirez-Soriano, B. Oliva, F. Calafell, J. Bertranpetit, and E. Bosch. 2008. "Signatures of selection in the human olfactory receptor OR5I1 gene." *Molecular Biology and Evolution* 25 (1):144-54.

Moreno-Estrada, A., C. R. Gignoux, J. C. Fernandez-Lopez, F. Zakharia, M. Sikora, A. V. Contreras, V. Acuna-Alonzo, K. Sandoval, C. Eng, S. Romero-Hidalgo, P. Ortiz-Tello, V. Robles, E. E. Kenny, I. Nuno-Arana, R. Barquera-Lozano, G. Macin-Perez, J. Granados-Arriola, S. Huntsman, J. M. Galanter, M. Via, J. G. Ford, R. Chapela, W. Rodriguez-Cintron, J. R. Rodriguez-Santana, I. Romieu, J. J. Sienra-Monge, B. del Rio Navarro, S. J. London, A. Ruiz-Linares, R. Garcia-Herrera, K. Estrada, A. Hidalgo-Miranda, G. Jimenez-Sanchez, A. Carnevale, X. Soberon, S. Canizales-Quinteros, H. Rangel-Villalobos, I. Silva-Zolezzi, E. G. Burchard, and C. D. Bustamante. 2014. "The genetics of Mexico recapitulates Native American substructure and affects biomedical traits." *Science* 344 (6189):1280-1285.

Moreno-Estrada, A., S. Gravel, F. Zakharia, J. L. McCauley, J. K. Byrnes, C. R. Gignoux, P. A. Ortiz-Tello, R. J. Martinez, D. J. Hedges, R. W. Morris, C. Eng, K. Sandoval, S. Acevedo-Acevedo, P. J. Norman, Z. Layrisse, P. Parham, J. C. Martínez-Cruzado, E. G. Burchard, M. L. Cuccaro, E. R. Martin, and C. D. Bustamante. 2013. "Reconstructing the population genetic history of the Caribbean." *PLoS Genetics* 9 (11):e1003925.

Morgan, K. 2007. *Slavery and the British Empire*. Oxford: Oxford University Press.

Morsink, J. 2013. "Exchange as a Social Contract: A Perspective from the Microscale." In *The Oxford Handbook of Caribbean Archaeology*, edited by W. F. Keegan, C. L. Hofman and R. Rodriguez Ramos, 312-328. Oxford: Oxford University Press.

Moya Pons, F. 1999. "The establishment of primary centres and primary plantations." In *General History of the Caribbean. Volume II: New Societies: The Caribbean in the Long Sixteenth Century*, edited by P. C. Emmer, 62-78. London: UNESCO Publishing.

Nei, M. 1987. *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nei, M., and J. Miller. 1990. "A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data." *Genetics* 125 (4):873-879.

Nelson, M. R., K. Bryc, K. S. King, A. Indap, A. R. Boyko, J. Novembre, L. P. Briley, Y. Maruyama, D. M. Waterworth, G. Waeber, P. Vollenweider, J. R. Oksenberg, S. L. Hauser, H. A. Stirnadel, J. S. Kooner, J. C. Chambers, B. Jones, V. Mooser, C. D. Bustamante, A. D. Roses, D. K. Burns, M. G. Ehm, and E. H. Lai. 2008. "The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research." *American Journal of Human Genetics* 83 (3):347-58.

Neuwirth, E. RColorBrewer: ColorBrewer Palettes. , R package version 1.1-2. https://cran.r-project.org/package=RColorBrewer.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. S. J, M. D. Adams, and M. Cargill. 2005. "A scan for positively selected genes in the genomes of humans and chimpanzees." *PLoS Biology* 3 (6):e170.

Nieves Colón, M. A., H. J. Díaz-Zabala, and J. C. Martínez-Cruzado. 2012. "The Contribution of sub-Saharan African and Eurasian maternal (mtDNA) lineages to the genetic heritage of the Dominican Republic." Master's Thesis M.A. Thesis, School of Human Evolution and Social Change, Arizona State University.

Nistal-Moret, B. 1984. *Esclavos, prófugos y cimarrones: Puerto Rico, 1770-1870*. Rio Piedras: Editorial de la Universidad de la Puerto Rico.

O'Rourke, D. H. 2012. "Why do we migrate? A retrospective." In *Causes and Consequences of Human Migration: An Evolutionary Perspective*, edited by M. H. Crawford and B. C. Campbell. Cambridge: Cambridge University Press.

O'Rourke, D. H., and J. A. Raff. 2010. "The human genetic history of the Americas: the final frontier." *Current Biology* 20 (4):R202-7.

Oboler, S. 1995. "Establishing an identity" in the sixties: The Mexican-American/Chicano and Puerto Rican movements." In *Ethnic Labels, Latino Lives: Identity and the Politics of (re)presentation in the United States*, 44-79. Minneapolis: University of Minnesota Press

Oh, S. S., J. Galanter, N. Thakur, M. Pino-Yanes, N. E. Barcelo, M. J. White, D. M. de Bruin, R. M. Greenblatt, K. Bibbins-Domingo, A. H. B. Wu, L. N. Borrell, C. Gunter, N. R. Powe, and E. G. Burchard. 2015. "Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled." *PLoS Medicine* 12 (12):e1001918.

Okonechnikov, K., A. Conesa, and F. García-Alcalde. 2016. "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." *Bioinformatics* 32 (2):292-294.

Oleksyk, T. K., M. W. Smith, and S. J. Brien. 2009. "Genome-wide scans for footprints of natural selection." *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1537):185.

Oliver, J. R. 2005. The Proto-Taíno Monumental Cemís of Caguana: A Political-Religious "Manifesto". In *Ancient Borinquen: Archaeology and Ethnohistory of Native Puerto Rico*, edited by P. Siegel. Tuscaloosa: University of Alabama Press.

Orlando, L., A. Ginolhac, G. Zhang, D. Froese, A. Albrechtsen, M. Stiller, M. Schubert, E. Cappellini, B. Petersen, I. Moltke, P. L. Johnson, M. Fumagalli, J. T. Vilstrup, M. Raghavan, T. Korneliussen, A. S. Malaspinas, J. Vogt, D. Szklarczyk, C. D. Kelstrup, J. Vinther, A. Dolocan, J. Stenderup, A. M. Velazquez, J. Cahill, M. Rasmussen, X. Wang, J. Min, G. D. Zazula, A. Seguin-Orlando, C. Mortensen, K. Magnussen, J. F. Thompson, J. Weinstock, K. Gregersen, K. H. Roed, V. Eisenmann, C. J. Rubin, D. C. Miller, D. F. Antczak, M. F. Bertelsen, S. Brunak, K. A. Al-Rasheid, O. Ryder, L. Andersson, J. Mundy, A. Krogh, M. T. Gilbert, K. Kjaer, T. Sicheritz-Ponten, L. J. Jensen, J. V. Olsen, M. Hofreiter, R. Nielsen, B. Shapiro, J. Wang, and E. Willerslev. 2013. "Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse." *Nature* 499 (7456):74-8.

Ousley, S. D., W. T. Billeck, and R. E. Hollinger. 2005. "Federal repatriation legislation and the role of physical anthropology in repatriation." *American Journal of Physical Anthropology* Suppl 41:2-32.

Overballe-Petersen, S., L. Orlando, and E. Willerslev. 2012. "Next-generation sequencing offers new insights into DNA degradation." *Annals of Anatomy* 30 (7):364-8.

Ozga, A. T., M. A. Nieves Colón, T. P. Honap, K. Sankaranarayanan, C. A. Hofman, G. R. Milner, C. M. Lewis, Jr., A. C. Stone, and C. Warinner. 2016. "Successful

enrichment and recovery of whole mitochondrial genomes from ancient human dental calculus." *American Journal of Physical Anthropology* 160 (2):220-8.

Pääbo, S. 1989. "Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification." *Proceedings of the National Academy of Sciences* 86:1939-1943.

Paijmans, J. L., M. T. Gilbert, and M. Hofreiter. 2013. "Mitogenomic analyses from ancient DNA." *Molecular Phylogenetics and Evolution* 69 (2):404-16.

Pantel, A. G. 2003. "The first Caribbean people: Part II: The Archaics." In *General History of the Caribbean Vol I: Autochtonous Societies*, edited by J. Sued-Badillo. London: UNESCO Publishing.

Paradis E., C. J., and K. Strimmer. 2004. "APE: analyses of phylogenetics and evolution in R language." *Bioinformatics* 20:289-290.

Parks, M., and D. Lambert. 2015. "Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study." *BMC Genomics* 16 (1):19.

Parra, E. J., R. A. Kittles, and M. D. Shriver. 2004. "Implications of correlations between skin color and genetic ancestry for biomedical research." *Nature Genetics* 36 (11 Suppl):S54-60.

Pasaniuc, B., S. Sankararaman, D. G. Torgerson, C. Gignoux, N. Zaitlen, C. Eng, W. Rodriguez-Cintron, R. Chapela, J. G. Ford, P. C. Avila, J. Rodriguez-Santana, G. K. Chen, L. Le Marchand, B. E. Henderson, D. Reich, C. Haiman, E. Gonzalez-Burchard, and E. Halperin. 2013. "Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation." *Bioinformatics* 29 (1):1407-1415.

Patterson, N., A. L. Price, and D. Reich. 2006. "Population structure and eigenanalysis." *PLoS Genetics* 2 (12):e190.

Peltzer, A., G. Jäger, A. Herbig, A. Seitz, C. Kniep, J. Krause, and K. Nieselt. 2016. "EAGER: efficient ancient genome reconstruction." *Genome Biology* 17 (1):60.

Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li. 2012. "Genomic patterns of homozygosity in worldwide human populations." *American Journal of Human Genetics* 91 (2):275-92.

Perego, U. A., N. Angerhofer, M. Pala, A. Olivieri, H. Lancioni, B. Hooshiar Kashani, V. Carossa, J. E. Ekins, A. Gomez-Carballa, G. Huber, B. Zimmermann, D. Corach, N. Babudri, F. Panara, N. M. Myres, W. Parson, O. Semino, A. Salas, S. R. Woodward, A. Achilli, and A. Torroni. 2010. "The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia." *Genome Research* 20 (9):1174-9.

Perego, U. A., H. Lancioni, M. Tribaldos, N. Angerhofer, J. E. Ekins, A. Olivieri, S. R. Woodward, J. M. Pascale, R. Cooke, J. Motta, and A. Achilli. 2012. "Decrypting the mitochondrial gene pool of modern Panamanians." *PLoS One* 7 (6):e38337.

Pérotin-Dumon, A. 1999. "French, English and Dutch in the Lesser Antilles: from privateering to planting,c. 1550-c.1650." In *General History of the Caribbean. Volume II: New Societies: The Caribbean in the Long Sixteenth Century*, edited by P. C. Emmer, 114-158. London: UNESCO Publishing.

Pestle, W. J. 2010. "Diet and Society in Prehistoric Puerto Rico: An Isotopic Approach." PhD, Graduate College University of Illinois at Chicago.

Pestle, W. J., and M. Colvard. 2012. "Bone collagen preservation in the tropics: a case study from ancient Puerto Rico." *Journal of Archaeological Science* 39 (7):2079-2090.

Petersen, J. B., C. L. Hofman, and L. A. Curet. 2004. "Time and culture: Chronology and taxonomy in the Eastern Caribbean and the Guianas." In *Late Ceramic Age Societies in the Eastern Caribbean*, edited by A. Delpuech and C. L. Hofman. Oxford: Archaeopress.

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, D. Absher, B. S. Srinivasan, G. S. Barsh, R. M. Myers, M. W. Feldman, and J. K. Pritchard. 2009. "Signals of recent positive selection in a worldwide sample of human populations." *Genome Research* 19 (5):826-37.

Pickrell, J. K., and D. Reich. 2014. "Toward a new history and geography of human genes informed by ancient DNA." *Trends Genet* 30 (9):377-89.

Pinhasi, R., D. Fernandes, K. Sirak, M. Novak, S. Connell, S. Alpaslan-Roodenberg, F. Gerritsen, V. Moiseyev, A. Gromov, P. Raczky, A. Anders, M. Pietrusewsky, G. Rollefson, M. Jovanovic, H. Trinhhoang, G. Bar-Oz, M. Oxenham, H. Matsumura, and M. Hofreiter. 2015. "Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone." *PLoS One* 10 (6):e0129102.

Popejoy, A. B., and S. M. Fullerton. 2016. "Genomics is failing on diversity." *Nature* 538:161-164.

Price, A., M. E. Weale, N. Patterson, S. R. Myers, A. C. Need, K. V. Shianna, D. Ge, J. I. Rotter, E. Torres, K. D. Taylor, D. B. Goldstein, and D. Reich. 2008. "Long-range LD can confound genome scans in admixed populations." *American Journal of Human Genetics* 83:127-147.

Prugnolle, F., A. Manica, M. Charpentier, J. F. Guegan, V. Guernier, and F. Balloux. 2005. "Pathogen-driven selection and worldwide HLA class I diversity." *Current Biology* 15 (11):1022-7.

Pugach, I., and M. Stoneking. 2015. "Genome-wide insights into the genetic history of human populations." *Investigative Genetics* 6:6.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3):559-575.

Purves, D., G. J. Augustine, D. Fitzpatrick, L. C. Katz, A.-S. LaMantia, J. O. McNamara, and S. M. Williams. 2001. *Neuroscience*. 2nd ed. USA: Sinauer Associates.

Pybus, M., G. M. Dall'Olio, P. Luisi, M. Uzkudun, A. Carreño-Torres, P. Pavlidis, H. Laayouni, J. Bertranpetit, and J. Engelken. 2013. "1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans." *Nucleic Acids Research*.

Quinlan, A. R., and I. M. Hall. 2010. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26 (6):841-842.

R Core Team. R: A language and environment for statistical computing. URL https://www.r-project.org/.

Racimo, F., G. Renaud, and M. Slatkin. 2016. "Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans." *PLoS Genetics* 12 (4):e1005972.

Raghavan, M., M. Steinrucken, K. Harris, S. Schiffels, S. Rasmussen, M. DeGiorgio, A. Albrechtsen, C. Valdiosera, M. C. Avila-Arcos, A. S. Malaspinas, A. Eriksson, I. Moltke, M. Metspalu, J. R. Homburger, J. Wall, O. E. Cornejo, J. V. Moreno-Mayar, T. S. Korneliussen, T. Pierre, M. Rasmussen, P. F. Campos, B. Damgaard Pde, M. E. Allentoft, J. Lindo, E. Metspalu, R. Rodriguez-Varela, J. Mansilla, C. Henrickson, A. Seguin-Orlando, H. Malmstrom, T. Stafford, Jr., S. S. Shringarpure, A. Moreno-Estrada, M. Karmin, K. Tambets, A. Bergstrom, Y. Xue, V. Warmuth, A. D. Friend, J. Singarayer, P. Valdes, F. Balloux, I. Leboreiro, J. L. Vera, H. Rangel-Villalobos, D. Pettener, D. Luiselli, L. G. Davis, E. Heyer, C. P. Zollikofer, M. S. Ponce de Leon, C. I. Smith, V. Grimes, K. A. Pike, M. Deal, B. T. Fuller, B. Arriaza, V. Standen, M. F. Luz, F. Ricaut, N. Guidon, L. Osipova, M. I. Voevoda, O. L. Posukh, O. Balanovsky, M. Lavryashina, Y. Bogunov, E. Khusnutdinova, M. Gubina, E. Balanovska, S. Fedorova, S. Litvinov, B. Malyarchuk, M. Derenko, M. J. Mosher, D. Archer, J. Cybulski, B. Petzelt, J. Mitchell, R. Worl, P. J. Norman, P. Parham, B. M. Kemp, T. Kivisild, C. Tyler-Smith, M. S. Sandhu, M. Crawford, R. Villems, D. G. Smith, M. R. Waters, T. Goebel, J. R. Johnson, R. S. Malhi, M. Jakobsson, D. J. Meltzer, A. Manica, R. Durbin, C. D. Bustamante, Y. S. Song, R. Nielsen, and E. Willerslev. 2015. "Genomic evidence for the Pleistocene and recent population history of Native Americans." *Science* 349 (6250):aab3884.

Ramos-Zayas, A. Y. 2003. *National Performances: The Politics of Class, Race and Space in Puerto Rican Chicago*. Chicago: University of Chicago Press.

Reich, D., N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M. V. Parra, W. Rojas, C. Duque, N. Mesa, L. F. Garcia, O. Triana, S. Blair, A. Maestre, J. C. Dib, C. M. Bravi, G. Bailliet, D. Corach, T. Hunemeier, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, V. Acuna-Alonzo, C. Aguilar-Salinas, S. Canizales-Quinteros, T. Tusie-Luna, L. Riba, M. Rodriguez-Cruz, M. Lopez-Alarcon, R. Coral-Vazquez, T. Canto-Cetina, I. Silva-Zolezzi, J. C. Fernandez-Lopez, A. V. Contreras, G. Jimenez-Sanchez, M. J. Gomez-Vazquez, J. Molina, A. Carracedo, A. Salas, C. Gallo, G. Poletti, D. B. Witonsky, G. Alkorta-Aranburu, R. I. Sukernik, L. Osipova, S. A. Fedorova, R. Vasquez, M. Villena, C. Moreau, R. Barrantes, D. Pauls, L. Excoffier, G. Bedoya, F. Rothhammer, J. M. Dugoujon, G. Larrouy, W. Klitz, D. Labuda, J. Kidd, K. Kidd, A. Di Rienzo, N. B. Freimer, A. L. Price, and A. Ruiz-Linares. 2012. "Reconstructing Native American population history." *Nature* 488 (7411):370-4.

Reich, D., N. Patterson, M. Kircher, F. Delfin, M. R. Nandineni, I. Pugach, A. M. Ko, Y. C. Ko, T. A. Jinam, M. E. Phipps, N. Saitou, A. Wollstein, M. Kayser, S. Paabo, and M. Stoneking. 2011. "Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania." *American Journal of Human Genetics* 89 (4):516-28.

Reid, B. A. 2014. The Archaic Age. In *Encyclopedia of Caribbean Archaeology*, edited by B. A. Reid and G. R. Gilmore III. Gainesville: University Press of Florida.

Richards, M. B., B. C. Sykes, and R. E. M. Hedges. 1995. "Authenticating DNA Extracted From Ancient Skeletal Remains." *Journal of Archaeological Science* 22 (2):291-299.

Richardson, B. C. 2004. "The migration experience." In *General History of the Caribbean. Volume V: The Caribbean in the Twentieth Century.*, edited by B. Brereton, 434-464. London: UNESCO Publishing.

Rishishwar, L., A. B. Conley, C. H. Wigington, L. Wang, A. Valderrama-Aguirre, and I. K. Jordan. 2015. "Ancestry, admixture and fitness in Colombian genomes." *Scientific Reports* 5:12376.

Rivera-Pagán, L. N. 2003. "Freedom and servitude: Indigenous slavery and the Spanish conquest of the Caribbean." In *General History of the Caribbean. Volume I: Autochtonous Societies.*, edited by J. Sued-Badillo, 316-362. London: UNESCO Publishing.

Rodríguez López, M. 1991. "Arqueología de Punta Candelero, Puerto Rico." Proceedings of the 13th International Congress for Caribbean Archaeology, Curaçao.

Rodriguez Ramos, R. 2010a. *Caribbean Archaeology and Ethnohistory: Rethinking Puerto Rican Precolonial History*. Tuscaloosa: University of Alabama Press.

Rodriguez Ramos, R. 2010b. "Rethinking time in Caribbean archaeology: The Puerto Rico case study." In *Island shores, distant pasts: Archaeological and biological approaches to the Pre-Columbian settlement of the Caribbean.*, edited by S. M. Fitzpatrick and A. H. Ross, 21-53. Gainesville: University Press of Florida.

Rodriguez Ramos, R. 2011. "Close Encounters of the Caribbean Kind." In *Caribbean Archaeology and Ethnohistory : Islands at the Crossroads : Migration, Seafaring,*

*and Interaction in the Caribbean*, edited by L. A. Curet and M. W. Hauser. Tuscaloosa: Unviersity of Alabama Press.

Rodriguez Ramos, R. 2013. "Isthmo-Antillean Engagements." In *The Oxford Handbook of Caribbean Archaeology*, edited by W. F. Keegan, C. L. Hofman and R. Rodríguez Ramos, 155-170. Oxford: Oxford University Press.

Rodriguez Ramos, R., J. Pagán Jiménez, and C. L. Hofman. 2013. "The humanization of the insular Caribbean." In *The Oxford Handbook of Caribbean Archaeology*, edited by W. F. Keegan, C. L. Hofman and R. Rodríguez Ramos, 126-140. Oxford: Oxford University Press.

Rogozinski, J. 2008. *A Brief History of the Caribbean: From the Arawak and Carib to the Present*. New York: Penguin Books.

Rohland, N., and M. Hofreiter. 2007. "Ancient DNA extraction from bones and teeth." *Nature Protocols* 2 (7):1756-62.

Rosenberg, N. A. 2006. "Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives." *Annals of Human Genetics* 70 (Pt 6):841-7.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. Feldman. 2002. "Genetic Structure of Human Populations." *Science* 298:2381-2385.

Ross, A. H. 2004. "Cranial evidence of pre-contact multiple population expansions in the Caribbean." *Caribbean Journal of Science* 40:291-298.

Ross, A. H., and D. H. Ubelaker. 2010. A morphometric approach to Taino biological distance in the Caribbean. In *Island shores, distant pasts: Archaeological and biological approaches to the Pre-Columbian settlement of the Caribbean*, edited by S. M. Fitzpatrick and A. H. Ross. Gainesville: University Press of Florida.

Ross, A. H., D. H. Ubelaker, and A. B. Falsetti. 2002. "Craniometric Variation in the Americas." *Human Biology* 74:807-818.

Rouse, I. 1986. "Migrations in Prehistory: Inferring Population Movement from Cultural Remains."

Rouse, I. 1992. *The Tainos: Rise & decline of the people who greeted Columbus*: Yale University Press.

Rousset, F. 2008. "genepop'007: a complete re-implementation of the genepop software for Windows and Linux." *Molecular Ecology Resources* 8 (1):103-106.

Ruaño, G., J. Duconge, A. Windemuth, C. L. Cadilla, M. Kocherla, D. Villagra, J. Renta, T. Holford, and P. J. Santiago-Borrero. 2009. "Physiogenomic analysis of the Puerto Rican population." *Pharmacogenomics* 10 (4):565-77.

Ruiz-Linares, A., K. Adhikari, V. Acuña-Alonzo, M. Quinto-Sanchez, C. Jaramillo, W. Arias, M. Fuentes, M. Pizarro, P. Everardo, F. de Avila, J. Gómez-Valdés, P. León-Mimila, T. Hunemeier, V. Ramallo, C. C. Silva de Cerqueira, M.-W. Burley, E. Konca, M. Z. de Oliveira, M. R. Veronez, M. Rubio-Codina, O. Attanasio, S. Gibbon, N. Ray, C. Gallo, G. Poletti, J. Rosique, L. Schuler-Faccini, F. M. Salzano, M.-C. Bortolini, S. Canizales-Quinteros, F. Rothhammer, G. Bedoya, D. Balding, and R. Gonzalez-José. 2014. "Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals." *PLoS Genetics* 10 (9):e1004572.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander, C. International HapMap, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I.

de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, T. A. Johnson, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart. 2007. "Genome-wide detection and characterization of positive selection in human populations." *Nature* 449 (7164):913-8.

Saint Pierre, A., and E. Génin. 2014. "How important are rare variants in common disease?" *Briefings in Functional Genomics*.

Salvaggio, J. E. 1992. "Fauna, flora, fowl and fruit: Effects of the Columbian exchange on the allergic response of New and Old World inhabitants." *Allergy and asthma proceedings* 13 (6):335-344.

Samaroo, B. 2011. "The immigrant communities." In *General History of the Caribbean. Volume IV: The Long Nineteenth Century: Nineteenth Century Transformations.*, edited by K. O. Laurence, 223-258. London: UNESCO Publishing.

Sankararaman, S., S. Mallick, M. Dannemann, K. Prufer, J. Kelso, S. Paabo, N. Patterson, and D. Reich. 2014. "The genomic landscape of Neanderthal ancestry in present-day humans." *Nature* 507 (7492):354-7.

Sans, M., G. Figueiro, C. E. Hughes, J. Lindo, P. C. Hidalgo, and R. S. Malhi. 2015. "A South American Prehistoric Mitogenome: Context, Continuity, and the Origin of Haplogroup C1d." *PLoS One* 10 (10):e0141808.

Sans, M., P. Mones, G. Figueiro, I. Barreto, J. M. Motti, M. D. Coble, C. M. Bravi, and P. C. Hidalgo. 2015. "The mitochondrial DNA history of a former native American village in northern Uruguay." *American Journal of Human Biology* 27 (3):407-16.

Scarano, F. A. 1992. *Haciendas y Barracones: Azucar y esclavitud en Ponce, Puerto Rico 1800-1850*. Rio Piedras: Ediciones Huracán.

Schlebusch, C. M., P. Skoglund, P. Sjodin, L. M. Gattepaille, D. Hernandez, F. Jay, S. Li, M. De Jongh, A. Singleton, M. G. Blum, H. Soodyall, and M. Jakobsson. 2012. "Genomic variation in seven Khoe-San groups reveals adaptation and complex African history." *Science* 338 (6105):374-9.

Schroeder, H., M. C. Ávila-Arcos, A.-S. Malaspinas, G. D. Poznik, M. Sandoval-Velasco, M. L. Carpenter, J. V. Moreno-Mayar, M. Sikora, P. L. F. Johnson, M. E. Allentoft, J. A. Samaniego, J. B. Haviser, M. W. Dee, T. W. Stafford, A. Salas, L. Orlando, E. Willerslev, C. D. Bustamante, and M. T. P. Gilbert. 2015. "Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean." *Proceedings of the National Academy of Sciences* 112 (12):3669-3673.

Schubert, M., L. Ermini, C. Der Sarkissian, H. Jonsson, A. Ginolhac, R. Schaefer, M. D. Martin, R. Fernandez, M. Kircher, M. McCue, E. Willerslev, and L. Orlando. 2014. "Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX." *Nature Protocols* 9 (5):1056-82.

Schuenemann, V. J., K. Bos, S. DeWitte, S. Schmedes, J. Jamieson, A. Mittnik, S. Forrest, B. K. Coombes, J. W. Wood, D. J. D. Earn, W. White, J. Krause, and H. N. Poinar. 2011. "Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death." *Proceedings of the National Academy of Sciences* 108 (38):E746–E752.

Schurr, T. G. 2004. "An anthropological genetic view of the peopling of the New World." In *The settlement of the American continents: a multidisciplinary approach to human biogeography*, edited by C. M. Barton, G. Clark, D. Yesner and G. Pearson. Arizona: University of Arizona Press.

Schurr, T. G. 2010. "Coastal Waves and Island Hopping: A genetic view of Caribbean prehistory and New World colonization." In *Island Shores, Distant Pasts: Archaeological and Biological approaches to the pre-Columbian settlement of the Caribbean.* , 177-198. Gainesville: University Press of Florida, Gainsville-Florida, USA.

Seguin-Orlando, A., C. A. Hoover, S. K. Vasiliev, N. D. Ovodov, B. Shapiro, A. Cooper, E. M. Rubin, E. Willerslev, and L. Orlando. 2015. "Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns." *STAR: Science & Technology of Archaeological Research* 1 (1):1-9.

Seguin-Orlando, A., T. S. Korneliussen, M. Sikora, A.-S. Malaspinas, A. Manica, I. Moltke, A. Albrechtsen, A. Ko, A. Margaryan, V. Moiseyev, T. Goebel, M. Westaway, D. Lambert, V. Khartanovich, J. D. Wall, P. R. Nigst, R. A. Foley, M. M. Lahr, R. Nielsen, L. Orlando, and E. Willerslev. 2014. "Genomic structure in Europeans dating back at least 36,200 years." *Science*.

Sheridan, R. B. 1972. "Africa and the Caribbean in the Atlantic Slave Trade." *The American Historical Review* 77 (1):15-35.

Siegel, P. 2005. *Ancient Borinquen: Archaeology and Ethnohistory of Native Puerto Rico*. Tuscaloosa: University of Alabama Press.

Siegel, P. 2011. "Puerto Rico." In *Protecting Heritage in the Caribbean*, edited by P. Siegel and E. Righter, 46-57. Tuscaloosa: University of Alabama Press.

Siegel, P. 2013. "Caribbean Archaeology in Historical Perspective." In *The Oxford Handbook of Caribbean Archaeology*, edited by W. F. Keegan, C. L. Hofman and R. Rodríguez Ramos, 21-46. Oxford: Oxford University Press.

Simbolo, M., M. Gottardi, V. Corbo, M. Fassan, A. Mafficini, G. Malpeli, R. T. Lawlor, and A. Scarpa. 2013. "DNA qualification workflow for next generation sequencing of histopathological samples." *PLoS One* 8 (6):e62692.

Simms, T. M., M. R. Wright, E. Martinez, M. Regueiro, Q. McCartney, and R. J. Herrera. 2013. "Y-STR diversity and sex-biased gene flow among Caribbean populations." *Gene* 516 (1):82-92.

Skoglund, P., S. Mallick, M. C. Bortolini, N. Chennagiri, T. Hunemeier, M. L. Petzl-Erler, F. M. Salzano, N. Patterson, and D. Reich. 2015. "Genetic evidence for two founding populations of the Americas." *Nature* 525 (7567):104-8.

Skoglund, P., H. Malmström, M. Raghavan, J. Storå, P. Hall, E. Willerslev, M. T. P. Gilbert, A. Götherström, and M. Jakobsson. 2012. "Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe." *Science* 336 (6080):466.

Skoglund, P., B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Paabo, J. Krause, and M. Jakobsson. 2014. "Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal." *Proc Natl Acad Sci U S A* 111 (6):2229-34.

Slatkin, M. 1985. "Gene flow in natural populations." *Annual Review of Ecology and Systematics* 16:393-430.

Slatkin, M., and F. Racimo. 2016. "Ancient DNA and human history." *Proc Natl Acad Sci U S A* 113 (23):6380-7.

Smith, C. I., A. T. Chamberlain, M. S. Riley, A. Cooper, C. B. Stringer, and M. J. Collins. 2001. "Not just old but old and cold?" *Nature* 410:771-772.

Soares, P., L. Ermini, N. Thomson, M. Mormina, T. Rito, A. Rohl, A. Salas, S. Oppenheimer, V. Macaulay, and M. B. Richards. 2009. "Correcting for purifying selection: an improved human mitochondrial molecular clock." *American Journal of Human Genetics* 84 (6):740-59.

Söchtig, J., V. Álvarez-Iglesias, A. Mosquera-Miguel, M. Gelabert-Besada, A. Gómez-Carballa, and A. Salas. 2015. "Genomic insights on the ethno-history of the Maya and the 'Ladinos' from Guatemala." *BMC Genomics* 16 (1):131.

Soler, M., M. Alves, M. S. Silva, M. Guimaraes, M. Sousa, J. Almeida, and E. Iwamura. 2011. "Morphological and DNA analysis in human skeletal remains exposed to environmental conditions in Brazil." *Forensic Science International Genetics Supplement* Series 3:339-340.

Steward, J. 1947. *The Circum-Caribbean Tribes*. Washington: Smithsonian Institution Bureau of American Ethnology.

Szpiech, Z. A., and R. D. Hernandez. 2014. "selscan: an efficient multithreaded program to perform EHH-based scans for positive selection." *Molecular Biology and Evolution* 31 (10):2824-7.

Tamm, E., T. Kivisild, M. Reidla, M. Metspalu, D. G. Smith, C. J. Mulligan, C. M. Bravi, O. Rickards, C. Martinez-Labarga, E. K. Khusnutdinova, S. A. Fedorova, M. V. Golubenko, V. A. Stepanov, M. A. Gubina, S. I. Zhadanov, L. P. Ossipova, L. Damba, M. I. Voevoda, J. E. Dipierri, R. Villems, and R. S. Malhi. 2007. "Beringian Standstill and Spread of Native American Founders." *PLoS One* 2 (9):e829.

Tamura, K., G. Stecher, D. Peterson, A. Filipski, and S. Kumar. 2013. "MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0." *Molecular Biology and Evolution* 30 (12):2725-2729.

Tang, H., S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Cintron, E. G. Burchard, and N. J. Risch. 2007. "Recent genetic selection in the ancestral admixture of Puerto Ricans." *American Journal of Human Genetics* 81 (3):626-33.

Tishkoff, S. 2015. "Strength in small numbers." *Science* 349 (6254):1282.

Toro-Labrador, G., O. R. Wever, and J. C. Martínez-Cruzado. 2003. "Mitochondrial DNA Analysis in Aruba: Strong Maternal Ancestry of Closely Related Amerindians and Implications for the Peopling of Northwestern Venezuela." *Caribbean Journal of Science* 39 (1):11-22.

Torroni, A., T. G. Schurr, M. F. Cabell, M. D. Brown, J. V. Neel, M. Larsen, D. G. Smith, C. M. Vullo, and D. C. Wallace. 1993. "Asian affinities and continental radiation of the four founding Native American mtDNAs." *American Journal of Human Genetics* 53 (3):563-90.

Turner, S. qqman: Q-Q and manhattan plots for GWAS data. R package version 0.1.2. https://cran.r-project.org/package=qqman.

Valdiosera, C., N. Garcia, L. Dalen, C. Smith, R. D. Kahlke, K. Liden, A. Angerbjorn, J. L. Arsuaga, and A. Gotherstrom. 2006. "Typing single polymorphic nucleotides in mitochondrial DNA as a way to access Middle Pleistocene DNA." *Biology Letters* 2 (4):601-3.

van Oven, M., and M. Kayser. 2009. "Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation." *Human Mutation* 30 (2):E386-E394.

Veran, C. 2003. "Born Puerto Rican, born (again) Taino? A resurgence of indigenous identity among Puerto Ricans has sparked debates over the island's tri-racial history." *Colorlines Magazine*, 23-25.

Via, M., C. R. Gignoux, L. A. Roth, L. Fejerman, J. Galanter, S. Choudhry, G. Toro-Labrador, J. Viera-Vera, T. K. Oleksyk, K. Beckman, E. Ziv, N. Risch, E. G. Burchard, and J. C. Martínez-Cruzado. 2011. "History shaped the geographic distribution of genomic admixture on the island of Puerto Rico." *PLoS One* 6 (1):e16513.

Vilar, M. G., C. Melendez, A. B. Sanders, A. Walia, J. B. Gaieski, A. C. Owings, T. G. Schurr, and C. Genographic. 2014. "Genetic diversity in Puerto Rico and its implications for the peopling of the Island and the West Indies." *American Journal of Physical Anthropology* 155 (3):352-68.

Vitti, J. J., S. R. Grossman, and P. C. Sabeti. 2013. "Detecting natural selection in genomic data." *Annual Reviews of Genetics* 47:97-120.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. "A map of recent positive selection in the human genome." *PLoS Biology* 4 (3):e72.

Wade, L. 2015. "Breaking a tropical taboo." *Science* 349 (6246):370.

Walker, J. 2005. "The Paso del Indio Site, Vega Baja, Puerto Rico: A Progress Report." In *Ancient Borinquen: Archaeology and Ethnohistory of Native Puerto Rico*, edited by P. Siegel, 55-87. Tuscaloosa: University of Alabama Press.

Walvin, J. 2013. *Crossings: Africa, the Americas and the Atlantic Slave Trade*. London: Reaktion Books.

Wang, S., N. Ray, W. Rojas, M. V. Parra, G. Bedoya, C. Gallo, G. Poletti, G. Mazzotti, K. Hill, A. M. Hurtado, B. Camrena, H. Nicolini, W. Klitz, R. Barrantes, J. A. Molina, N. B. Freimer, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, L. T. Tsuneto, J. E. Dipierri, E. L. Alfaro, G. Bailliet, N. O. Bianchi, E. Llop, F.

Rothhammer, L. Excoffier, and A. Ruiz-Linares. 2008. "Geographic patterns of genome admixture in Latin American Mestizos." *PLoS Genetics* 4 (3):e1000037.

Warinner, C., J. F. Rodrigues, R. Vyas, C. Trachsel, N. Shved, J. Grossmann, A. Radini, Y. Hancock, R. Y. Tito, S. Fiddyment, C. Speller, J. Hendy, S. Charlton, H. U. Luder, D. C. Salazar-Garcia, E. Eppler, R. Seiler, L. H. Hansen, J. A. Castruita, S. Barkow-Oesterreicher, K. Y. Teoh, C. D. Kelstrup, J. V. Olsen, P. Nanni, T. Kawai, E. Willerslev, C. von Mering, C. M. Lewis, Jr., M. J. Collins, M. T. Gilbert, F. Ruhli, and E. Cappellini. 2014. "Pathogens and host immunity in the ancient human oral cavity." *Nature Genetics* 46 (4):336-44.

Wegmann, D., D. E. Kessner, K. R. Veeramah, R. A. Mathias, D. L. Nicolae, L. R. Yanek, Y. V. Sun, D. G. Torgerson, N. Rafaels, T. Mosley, L. C. Becker, I. Ruczinski, T. H. Beaty, S. L. Kardia, D. A. Meyers, K. C. Barnes, D. M. Becker, N. B. Freimer, and J. Novembre. 2011. "Recombination rates in admixed individuals identified by ancestry-based inference." *Nature Genetics* 43 (9):847-53.

Weissensteiner, H., D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Specht, H.-J. Bandelt, F. Kronenberg, A. Salas, and S. Schönherr. 2016. "HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing." *Nucleic Acids Research* 44 (W1):W58-W63.

Wells, J. C. K., and J. T. Stock. 2012. "The biology of human migration: The ape that won't commit?" In *Causes and Consequences of Human Migration: An Evolutionary Perspective*, edited by M. H. Crawford and B. C. Campbell. Cambridge: Cambridge University Press.

Whitehead, N. L. 1999. "Native society and the European occupation of the Caribbean islands and coastal *Tierra Firme*, 1492-1650." In *General History of the Caribbean, Vol. II: New societies: The Caribbean in the long sixteenth century*, edited by P. C. Emmer, 180-201. London: UNESCO Publishing.

Wickham, H. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12):20.

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

Wickham, H. scales: Scale Functions for Visualization., R package version 0.4.1. https://cran.r-project.org/package=scales.

Wickham, H. tidyr: Easily Tidy Data with `spread()` and `gather()` Functions. R package version 0.6.0.

Williams, S. R., N. A. Chagnon, and R. S. Spielman. 2002. "Nuclear and mitochondrial genetic variation in the Yanomamo: a test case for ancient DNA studies of prehistoric populations." *American Journal of Physical Anthropology* 117 (3):246-59.

Williamson, S. H., M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante, and R. Nielsen. 2007. "Localizing recent adaptive evolution in the human genome." *PLoS Genetics* 3 (6):e90.

Wilson, S. M. 1999. "Cultural pluralism and the emergence of complex society in the Greater Antilles." Proceedings of the 18th International Congress for Caribbean Archaeology, Grenada.

Wilson, S. M. 2007. *The archaeology of the Caribbean*. Cambridge: Cambridge University Press.

Wollstein, A., and O. Lao. 2015. "Detecting individual ancestry in the human genome." *Investigative Genetics* 6:7.

Wright, S. 1951. "The genetical structure of populations." *Annals of Eugenics* 15:323-354.

Wright, S. 1978. *Evolution and the Genetics of Populations. Vol. 4. Variability Within and Among Natural Populations.* Chicago: University of Chicago Press.

Wu, D. D., D. M. Irwin, and Y. P. Zhang. 2008. "Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair." *BMC Evolutionary Biology* 8:241.

Yang, D. Y., B. Eng, J. S. Waye, J. C. Dudar, and S. R. Saunders. 1998. "Technical note: improved DNA extraction from ancient bones using silica-based spin columns." *American Journal of Physical Anthropology* 105:539-543.

Yates, A., W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino, and P. Flicek. 2015. "Ensembl 2016." *Nucleic Acids Research* 44 (D1):D710-D716.

Younger, R. M., C. Amadou, G. Bethel, A. Ehlers, K. F. Lindahl, S. Forbes, R. Horton, S. Milne, A. J. Mungall, J. Trowsdale, A. Volz, A. Ziegler, and S. Beck. 2001. "Characterization of clustered MHC-linked olfactory receptor genes in human and mouse." *Genome Research* 11 (4):519-30.

Zar, J. H. 2010. *Biostatistical Analysis*. Fifth ed. New Jersey: Pearson.

Zhou, Q., L. Zhao, and Y. Guan. 2016. "Strong Selection at MHC in Mexicans since Admixture." *PLoS Genetics* 12 (2):e1005847.

Ziegler, A. 1997. "Biology of Chromosome 6." *DNA Sequence* 8 (3):189-201.

Zong, J., and J. Batalova. 2016. Caribbean immigrants in the United States.

# APPENDIX A

# INSTITUTIONAL REVIEW BOARD APPROVALS AND MATERIAL TRANSFER AGREEMENTS FOR SAMPLES IN CHAPTER 2

Consult Attached Files using Adobe Acrobat Reader

# APPENDIX B

## SUPPLEMENTARY TABLES FOR CHAPTER 2 (SM1 TO SM3)

Consult Attached Files using Microsoft Excel

# APPENDIX C

## SUPPLEMENTARY FIGURES FOR CHAPTER 2 (SM1 TO SM14)

Consult Attached Files using Adobe Acrobat Reader

# APPENDIX D

# PROTOCOL FOR EXTRACTION METHOD H

**Protocol for Ancient DNA Extraction from Bone or Teeth (Method H)**

The following protocol isolates and purifies DNA from ancient bones or teeth. This method combines steps of previously published ancient DNA extraction protocols, with modifications. Total time: 2 days.

**Reagents**

0.5 M EDTA pH 8.0 (Life Technologies, cat.no. AM9262)

1 M Tris-HCl pH 8.0 (Life Technologies, cat.no. 15568-025)

Proteinase K (Qiagen, cat.no. 19133)

Guanidine hydrochloride (Sigma, cat.no. G3272)

3M sodium acetate pH 5.2 (Sigma, cat.no. S7899)

MinElute PCR purification kit (Qiagen, cat.no. 28004)

N-laurylsarcosyl 20% sodium salt solution (Sigma, cat.no. L7414)

Isopropanol (VWR, cat.no. BDH1133-4LP)

Nuclease free water ($ddH_20$) (Bioexpress, cat.no. G-3223-1L)

Tween 20 (Sigma, cat.no. P9416-50ML)

**Consumables**

1.5 mL LoBind tubes (Eppendorf, cat.no. 0030108051)

2.0 mL LoBind tubes (Eppendorf, cat.no. 0030108078)

50 mL Falcon tubes (VWR, cat.no. 352098)

15 mL Falcon tubes (VWR, cat.no. 352097)

Zymo-Spin V-Spin columns (ZymoResearch, cat. no. C1012-50) Treat with bleach, rinse in water and irradiate with UV before usage.

**Buffers**

Lysis buffer: 2 mL per sample (Must be made fresh, irradiate with UV before usage)

| Reagent | Final concentration | For 1 sample (2 mL) | For 7 samples (14 mL) |
|---|---|---|---|
| N-laurylsarcosyl 20% | 0.5% | 50 uL | 350 uL |
| 0.5 M EDTA, pH 8.0 | 0.45 M | 1.8 mL | 12.6 mL |
| 20 mg/ml Proteinase K | 0.25 mg/mL | 30 uL | 210 uL |
| $ddH_20$ | - | 130 uL | 910 uL |

TET buffer: 50 uL per sample, recipe for 10mL (Irradiate with UV before usage)

| Reagent | Final concentration | Volume |
|---|---|---|
| $ddH_20$ | n/a | 9.88 mL |
| 0.5 M EDTA, pH 8.0 | 1 mM | 20 uL |
| 1 M Tris-HCl, pH 8.0 | 10 mM | 100 uL |
| Tween 20 | 0.05% | 10 uL |

Binding buffer: 13 mL per sample, recipe for 15 mL to account for potential spillage (Can be made in advance on day 1, irradiate with UV before usage)

| Reagent | Final concentration | For 1 sample (15 mL) | For 7 samples (100 mL) |
|---|---|---|---|
| Guanidine hydrochloride | 5 M | 7.2 g | 47.7 g |
| $ddH_20$ | - | To 9 mL | To 60 mL |
| 100% Isopropanol | 40% | To 14.9 mL | To 90.9 mL |
| 100% Tween 20 | 0.05% | 10 uL | 50 uL |
| 3 M Sodium Acetate pH 5.2 | 90 mM | 450 uL | 3 mL |

Add salt first then add water (using the graduation of a falcon tube or Nalgene container). Mix to dissolve the salt (if necessary heat it in the microwave very briefly). Add isopropanol, then Tween 20 and Sodium Acetate. Making 15 mL for one sample to account for possible error or spillage.

**I. Sampling and Pre-Digestion to remove loose surface contaminants [EDTA Wash step] (Day 1: ~1-3 hours depending on samples) From: Warinner et al. (2014)**

1. Prepare three 2 mL SafeLock LoBind tubes per sample. Label one EDTA-WASH, one DIGEST, and one BONE PELLET. UV each tube for 10 minutes
2. For each sample, measure 100 mg of bone or dentine powder in BONE PELLET tube.
3. Add 1 mL of 0.5M EDTA to each sample. Seal caps tightly.
4. Vortex, then incubate on rotation at RT (~23°C) for 15 minutes. <u>Do not leave for longer.</u>
5. Centrifuge at 13,000 rpm for 1-3 minutes.
6. Transfer supernatant to EDTA-WASH tube. Save this tube in -20°C freezer until extraction process is completed.

**II. Decalcification and Digestion [Lysis Buffer step] (Day 1: ~2 hours) – From: Gamba et al. (2016)**

1. Add 1 mL of Lysis Buffer to sample pellet. Mix by pipetting up and down or vortexing.
2. Incubate on rotation for 1 hour at 56°C.
3. Centrifuge at 13,000 rpm for 2 minutes.
4. Collect supernatant into DIGEST tube. Save in 4°C fridge.
5. Add 1 mL of fresh Lysis Buffer to sample pellet. Incubate on rotation for an additional hour at 56°C and then overnight at 37°C.
6. Immerse one Zymo reservoir per sample in 50% bleach solution overnight.

**III. Binding and Purification (Day 2: ~3 hours) – From: Dabney et al (2013)**

1. Label one 15 mL Falcon tube, one 50 mL Falcon tube, one MinElute column, and one 1.5 mL SafeLock LoBind tube (for final elution) per sample.
2. Take Zymo reservoir out of bleach solution and wash with Sigma water. Dry with paper towel or Kimwipes.
3. UV all tubes and reservoirs for 10 minutes. Do not UV MinElute column.
4. Jam the Zymo reservoir into the MinElute column and place in pre-labeled 50 mL tubes.
5. Heat TET buffer to 65°C on heating block
6. For each sample, transfer 13 mL Binding Buffer to a 15 mL Falcon tube.
7. Centrifuge pellet sample tubes at 13,200 rpm (max speed) for 2 minutes.
8. Collect supernatant and add it to DIGEST tube. This combined digest will be used in further extraction procedures. Save sample pellet tube in -20°C freezer. It can be used in future extractions if necessary.
9. Transfer the DIGEST supernatant (2 mL) to falcon tube containing the Binding Buffer. Mix gently by shaking and let it stand for 10 minutes.
10. Pour the sample/Binding buffer mixture from into the extension reservoir and close the falcon tube with a screw cap. <u>Let stand for 5 minutes.</u>
11. Centrifuge for 4 min at 1,500 rpm in a large centrifuge. Turn the tubes by 180 degrees and centrifuge for 2 more min at 1,500 rpm.
12. Remove the screw cap and place the extension reservoir/spin column assembly back into the collection tube. Carefully remove the extension reservoir and close the cap of the spin column. Close the falcon tube with a screw cap and save the flow

through for later experiments.

13. Perform a dry spin for 1 minute at 6,000 rpm in a centrifuge. Discard the flow-through.
14. Add 750 uL PE buffer, spin at 6,000 rpm for 1 minute. Again, add 750 uL of PE buffer, and with the same tip, lift column and remove waste from collection tube.
15. Spin at 6,000 rpm for 1 minute again.
16. Remove waste from collection tube.
17. Perform a dry spin for 1 min at 13,200 rpm, turn the column 180 degrees and spin for 30 seconds. Transfer the spin column into a Lo-Bind (or siliconized) 1.5 ml final elution tube.
18. Add 25 uL TET (preheated to 65°C) on top of the silica membrane and let the tube stand for 10 minutes at RT. Spin for 1 minute at 13,200 rpm. Repeat this step. Ending volume of eluate: 50 uL
19. Transfer the eluate (the DNA extract) to a fresh tube. Aliquot 1 uL for Qubit quantification. Freeze remaining 49 uL at -20 °C.

# APPENDIX E

# DAMAGE PLOTS FOR SHOTGUN LIBRARIES IN CHAPTER 3

Consult Attached Files using Adobe Acrobat Reader

# APPENDIX F

# DAMAGE PLOTS FOR MTDNA ENRICHED LIBRARIES IN CHAPTER 3

Consult Attached Files using Adobe Acrobat Reader

# APPENDIX G

## SUPPLEMENTARY TABLES FOR CHAPTER 3 (SM4 TO SM6)

Consult Attached Files using Microsoft Excel

# APPENDIX H

## SUPPLEMENTARY FIGUES FOR CHAPTER 3 (SM15 TO SM29)

Consult Attached Files using Adobe Acrobat Reader

# APPENDIX I

## TABLE OF MAPPING STATISTICS AND HAPLOGROUP ASSIGNMENTS FOR 27 ANCIENT SAMPLES ANALYSED IN CHAPTER 4

**Mapping statistics and haplogroup assignments for 27 ancient samples with >3X read depth and 93% coverage of the mitochondrial genome**

Haplogroup assignments were originally performed in Haplogrep. Diagnostic positions were then revised manually. Scores listed in Table 1 correspond to second round of Haplogrep assignment on manually curated variants.

| Sample | Site | Radiocarbon date (median CalAD) | Total unique mtDNA reads | Mapped reads average length | Mean read depth (X) | StdDev read depth (X) | % ref-seq covered >1x | Haplogroup | Haplogroup quality score | Proportion Authentic [3] |
|---|---|---|---|---|---|---|---|---|---|---|
| PC-112 | Punta Candelero | 943 | 6,176 | 68.40 | 25 | 7.5 | 99.97% | C1b2 | 0.985 | 0.914 |
| PC-115 | Punta Candelero | None | 1,192 | 49.25 | 3.5 | 2.1 | 95.9% | C1b2 | 0.953 | 0.997 |
| PC-121[1] | Punta Candelero | 672 | 1,004 | 78.35 | 4.7 | 2.4 | 97.64% | C1b | 0.963 | 0.812 |
| PC-443[2] | Punta Candelero | 645 | 1,810 | 68.06 | 7.4 | 3.1 | 99.14% | C1b2 | 0.971 | 0.937 |
| PC-452 | Punta Candelero | 491 | 2,235 | 58.17 | 7.8 | 3.5 | 99.67% | D1 | 0.965 | 0.932 |
| PC-455 | Punta Candelero | 619 | 1,076 | 57.30 | 3.7 | 2.2 | 94.9% | A2 | 0.974 | 0.958 |
| PC-B4E2 | Punta Candelero | None | 4,091 | 56.53 | 13.9 | 4.9 | 99.86% | C1b2 | 0.986 | 0.987 |
| PC-E24[1] | Punta Candelero | None | 1,353 | 51.67 | 4.2 | 2.6 | 96.92% | C1b2 | 0.977 | 0.957 |
| PC-E54 | Punta Candelero | None | 1,904 | 48.00 | 5.5 | 2.8 | 99.29% | D1 | 0.964 | |
| **Average** | **Punta Candelero** | **674** | **2,316** | **59.5** | **8.4** | **3.4** | **98.42%** | **N/A** | **0.97** | **0.937** |
| PI-174 | Paso del Indio | 824 | 2,167 | 67.16 | 8.7 | 4.4 | 99.70% | C1b2 | 0.986 | 0.983 |
| PI-388[1] | Paso del Indio | 866 | 2,332 | 54.48 | 7.6 | 4.4 | 99.54% | C1b2 | 0.986 | 0.903 |
| PI-390 | Paso del Indio | 1167 | 5,765 | 63.35 | 22 | 7.7 | 99.95% | D1 | 0.957 | 0.857 |
| PI-395 | Paso del Indio | 901 | 7,801 | 57.62 | 27.1 | 9.8 | 99.99% | C1b2 | 0.986 | 0.940 |
| PI-397 | Paso del Indio | 1276 | 5,320 | 59.54 | 19.1 | 8.0 | 99.9% | C1b2 | 0.986 | 0.876 |
| PI-399 | Paso del Indio | 1095 | 11,372 | 69.80 | 47.9 | 12.4 | 99.98% | A2+(64) | 0.925 | 0.978 |
| PI-410 | Paso del Indio | None | 2,463 | 56.07 | 8.3 | 4.0 | 99.40% | A2e | 0.947 | 0.986 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| PI-417 | Paso del Indio | 1233 | 5,594 | 64.34 | 21.7 | 7.6 | 99.94% | C1b2 | 0.986 | 0.947 |
| PI-419 | Paso del Indio | 1122 | 18,068 | 65.04 | 70.9 | 13.7 | 100% | C1d1 | 0.897 | 0.978 |
| PI-420[1] | Paso del Indio | 1270 | 3,278 | 60.97 | 12 | 5.2 | 99.89% | A2 | 0.957 | 0.970 |
| PI-423 | Paso del Indio | 965 | 4,478 | 54.13 | 14.6 | 6.3 | 99.80% | A2e | 0.922 | 0.978 |
| PI-424 | Paso del Indio | None | 2,449 | 65.66 | 9.7 | 4.2 | 99.61% | C1b2 | 0.962 | 0.978 |
| PI-435 | Paso del Indio | 994 | 909 | 67.53 | 3.7 | 2.1 | 95.56% | A2e | 0.942 | 0.923 |
| PI-437 | Paso del Indio | 1039 | 1,320 | 65.88 | 5.2 | 2.7 | 97.80% | A2e | 0.905 | 0.731 |
| PI-48[1] | Paso del Indio | 674 | 4,089 | 61.04 | 15 | 7.4 | 99.97% | C1b2 | 0.986 | 0.931 |
| PI-67[1] | Paso del Indio | 1022 | 14,599 | 62.01 | 54.6 | 16.1 | 99.99% | A2+(64) | 0.922 | 0.959 |
| **Average** | **Paso del Indio** | **1032** | **5,750** | **62.2** | **21.8** | **7.3** | **99.14%** | **N/A** | **0.95** | **0.932** |
| T-251[1] | Tibes | 616 | 4,830 | 62.26 | 18.1 | 6.1 | 99.95% | D1 | 0.895 | 0.958 |
| T-257 | Tibes | 878 | 15,413 | 72.99 | 67.9 | 14.4 | 99.99% | C1c | 0.967 | 0.974 |
| **Average** | **Tibes** | **747** | **10121.5** | **67.63** | **43.0** | **10.2** | **99.00%** | **N/A** | **0.93** | **0.966** |
| **ALL** | **ALL** | **920** | **4929** | **61.7** | **18.9** | **6.2** | **98.99%** | **N/A** | **0.96** | **0.936** |

[1] Reads combined across multiple sequencing runs.

[2] Reads subjected to contamination filtering with PMDTools.

[3] Estimated proportion of authentic endogenous human reads using contamMix.

# APPENDIX J

## SUPPLEMENTARY TABLES FOR CHAPTER 4 (SM7 TO SM22)

Consult Attached Files using Microsoft Excel

# APPENDIX K

## SUPPLEMENTARY TABLES FOR CHAPTER 4 (SM30 TO SM34)

Consult Attached Files using Adobe Acrobat Reader