INVESTIGATING GENES RELATED TO THE EVOLUTION OF
INDIGENOUS AMERICAN SKIN PIGMENTATION

A Dissertation in
Anthropology

by

Ellen E. Quillen

© 2010 Ellen E. Quillen

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2010
The dissertation of Ellen Quillen was reviewed and approved* by the following:

Mark D. Shriver  
Associate Professor of Biological Anthropology  
Dissertation Advisor  
Chair of Committee

Kenneth M. Weiss  
Evan Pugh Professor of Biological Anthropology and Genetics and Science,  
Technology, and Society

David A. Puts  
Assistant Professor of Biological Anthropology

David J. Vandenergh  
Associate Professor of Biobehavioral Health

Nina G. Jablonski  
Professor of Anthropology  
Head of the Department of Anthropology

*Signatures are on file in the Graduate School
ABSTRACT

The genetic study of human pigmentation is fundamental to anthropology because pigmentation represents an enormously variable phenotype present in all humans shaped and reshaped over the history of the hominid lineage by the forces of evolution. Skin pigmentation is among the most conspicuously varying traits among individuals and populations and must be understood in the context of the evolutionary history of the human species. As a population moves into a new environment, natural selection acts on both existing and newly emerging genetic variation in the population. There is a clear north-south gradient of skin pigmentation across the globe, with populations at higher latitudes having less melanin in their skin on average than populations at lower latitudes. This distribution suggests that skin pigmentation has been under strong selection as human populations spread throughout Africa and eventually around the world. There are a number of hypotheses to explain why this distribution occurs, predominantly rooted in the distribution of ultra-violet radiation (UVR). Previous researchers have employed statistical methods which detect distributions of allele frequencies and genomic signatures inconsistent with neutral evolution. These methods have identified skin pigmentation genes in a number of Old World populations that appear to have undergone selection. This dissertation reports on the first genome-wide analysis of evidence for selection at skin pigmentation genes in New World populations.

From 76 initial pigmentation candidate genes, four tests of selection – Locus-Specific Branch Length (LSBL), Log of the Ratio of Heterozygosities (lnRH), Tajima’s D, and a heuristic haplotypes analysis – were used to identify fourteen pigmentation genes with signatures of selection unique to Indigenous American populations. Based on the evidence of selection, a logical hypothesis is these genes may play a role in determining skin color in Indigenous American populations. To test this hypothesis, admixture linkage analysis was performed in
several samples of European and Indigenous American ancestry. Admixture linkage analysis exploits a number of unique properties of admixed populations. In an admixed population, genes influencing skin pigmentation variation between the Indigenous American and European parental populations will have a higher proportion of Indigenous American alleles than expected in individuals with darker skin pigmentation. Most importantly, the use of admixed populations allows for the appearance of heterozygotes at loci where alternate alleles may be fixed in the parental populations.

To characterize these admixed populations, ancestry informative markers (AIMs) were genotyped in the samples. In particular, the distribution of biogeographic ancestry as described by both nuclear AIMs and male-specific non-recombinant Y chromosome markers is discussed for a sample of 173 Colombian individuals. The distribution of biogeographic ancestry (BGA) in this population includes ancestors from West Africa, America, and Western Europe. Although these individuals self-report ethnic identifications that are generally consistent with their BGA, the distribution of BGA among the ethnic groups shows a large degree of overlap which demonstrates the heterogeneous nature of these ethnic groups. Additionally, the Y-chromosomal markers, which are inherited exclusively through the male line can be traced to a likely geographic region of origin. The ancestry estimates from the nuclear and Y-chromosomal markers indicate that there is a bias towards more European males and more African and Indigenous American females contributing to the gene pool. This sex-biased gene flow is common in admixed populations of the Americas.

In total, 515 individuals with mixed Indigenous American and European ancestry were genotyped for single nucleotide polymorphisms (SNPs) chosen to be ancestry informative between the European and Indigenous American parental populations. These SNPs were used to test for admixture linkage association between the candidate genes and variation in skin color. The results of this analysis show a novel association of the genes PAX3 and ASIP with mean
constitutive skin pigmentation differences between European and Indigenous American populations. The roles of SLC24A5 and MATP (SLC45A2) in contributing to skin pigmentation differences between European and non-European populations were also confirmed. SLC24A5 and MATP (SLC45A2) are each linked to approximately 3 melanin units of variation in skin pigmentation while PAX3 and ASIP each contribute to less than 1.5 melanin units of difference.
TABLE OF CONTENTS

List of Figures .................................................................................................................. ix

List of Tables ..................................................................................................................... xi

Acknowledgements .......................................................................................................... xii

Chapter 1 Introduction and Background ........................................................................ 1

Importance of the Study of Human Skin Pigmentation ....................................................... 4

Worldwide Distribution of Human Skin Pigmentation ......................................................... 6

Hypotheses for Selective Pressures on Skin Pigmentation .................................................. 10

Ultraviolet Radiation and Skin Cancer ............................................................................. 10

Ultraviolet Radiation and Folic Acid Photolysis ................................................................. 11

Alternate Theories Favoring Dark Skin Pigmentation ......................................................... 12

Ultraviolet Radiation and Vitamin D Production ............................................................... 13

Sexual Selection in Skin Pigmentation .............................................................................. 15

Genes Involved in Melanogenesis ..................................................................................... 17

Skin Pigmentation in the New World ................................................................................. 20

References .......................................................................................................................... 23

Chapter 2 Evidence for Natural Selection at Pigmentation Genes in the Americas .......... 27

Methods for Identifying Selection in the Genome ............................................................... 28

Selection at Skin Pigmentation Genes in Old World Populations ...................................... 33

Populations and Genotyping Platforms ............................................................................. 37

Mega Samples .................................................................................................................... 38

Human Genome Diversity Project Samples ....................................................................... 42

Testing for Selection in Indigenous American Populations ................................................. 44

Locus Specific Branch Length ........................................................................................... 46

Natural Log Ratio of Heterozygosities (lnRH) .................................................................... 51

Tajima’s D ............................................................................................................................ 52

Assessment of Haplotype Structure .................................................................................. 56

Necessity of Incorporating Multiple Tests ........................................................................ 57

Power to Detect Selection ................................................................................................. 59

Results of Tests for Selection ............................................................................................ 62

Identification of Selection-Nominated Candidate Genes .................................................. 67

Strength of Selection ......................................................................................................... 85

Conclusion .......................................................................................................................... 86

References .......................................................................................................................... 87

Chapter 3 Admixture, Ethnicity, and Skin Pigmentation in Southwestern Colombia ........ 91

Limitations and Merits of Different Types of AIMs ........................................................... 92

Uniparental Markers .......................................................................................................... 93

Autosomal Markers ........................................................................................................... 95
Admixture Studies in South America ......................................................... 96
  Sex-Biased Gene Flow ........................................................................... 97
  Admixture in Colombia ......................................................................... 98
  Admixture in Neighboring Countries ..................................................... 101
Chapter 4 Conclusion ........................................................................... 104
  Samples and Methods ........................................................................ 104
    Description of Samples ..................................................................... 104
    Y-Chromosome Genotyping ............................................................... 105
    Autosomal AIMS Genotyping ............................................................. 106
  Methods............................................................................................... 107
    Y-Chromosomal Ancestry Estimates .................................................. 107
    Autosomal Ancestry Estimates ......................................................... 112
    Ancestry, Ethnicity, and Phenotypes ................................................ 116
  Results and Discussion .................................................................... 119
    Genotype Distributions .................................................................... 148
    ADMIXMAP Results ........................................................................ 148
    Power Calculations ........................................................................ 151
    Kruskal-Wallis Tests ....................................................................... 158
    Mann-Whitney U Tests .................................................................... 161
  Conclusions ...................................................................................... 164
  References ...................................................................................... 165

Chapter 5 Conclusion ........................................................................... 167
  Relationship between Biogeographic Ancestry and Ethnicity ............... 167
  Genes Contributing to Skin Pigmentation Variation .............................. 169
    PAX3 ............................................................................................... 169
    ASIP ............................................................................................... 170
    SLC24A5 and MATP ........................................................................ 171
  Necessity of Testing Functionality for Signatures of Selection ............... 171
  Future Directions ............................................................................. 173
    Genes for Future Investigation ......................................................... 173
    Dating Rise of Mutations ................................................................. 174
  References ...................................................................................... 177

Appendix A  PERL Scripts for Calculating Test Statistics ....................... 179
  LSBL ................................................................................................. 179
  LnRH ............................................................................................... 182
  Tajima’s D ....................................................................................... 184
Tajima’s D Difference ........................................................................................................189

Appendix B  Results of Selection Analysis ..................................................................... 192

Appendix C  Parental Population Frequencies ............................................................269

Burchard AIMs Panel (Used for Colombian Sample) ................................................. 269
Bonilla AIMs Panel (Used for San Luis Valley and Guerrero Samples) ................. 271
DNAPrint AIMs Panel (Used for UNM) .................................................................. 272
LIST OF FIGURES

Figure 1-1. Distribution of global skin pigmentation ........................................... 7
Figure 1-2. Distribution of ultraviolet radiation and skin pigmentation world-wide .......... 9
Figure 1-3. Melanosome distribution in three populations ........................................ 18
Figure 1-4. Graphical representation of a melanocyte ............................................. 19
Figure 2-1. Effects of natural selection on gene trees ................................................. 29
Figure 2-2. Informative range of tests of selection relative to human evolution .................. 32
Figure 2-3. Location of samples used in selection analyses ......................................... 41
Figure 2-4. Diagram and formula for calculating LSBL .............................................. 49
Figure 2-5. Estimated False Discovery or False Negative Rates using Empirical Models .... 61
Figure 2-6. Graphical representations of evidence for selection in selection-nominated candidate genes ................................................................. 82
Figure 3-1. Inheritance of uniparental and biparental markers ..................................... 93
Figure 3-2. Political map of the departments of Colombia ......................................... 99
Figure 3-3. Diagnostic SNPs for Y haplogroups in sample ........................................... 106
Figure 3-4. Distribution of Y haplogroups by ethnicity ............................................. 108
Figure 3-5. World-wide distribution among indigenous groups of most common Colombian haplotypes ................................................................. 109
Figure 3-6. Y-chromosome origins by ethnicity ......................................................... 111
Figure 3-7. Distribution of autosomal ancestry estimates by ethnicity ........................... 114
Figure 3-8. Distribution of Melanin Index by ethnicity ............................................... 117
Figure 3-9. Distribution of ancestry-adjusted residual melanin by ethnicity .................... 118
Figure 4-1. Effects of admixture in a population over multiple generations ..................... 128
Figure 4-2. Interrelationships between parameters estimated by ADMIXMAP .................. 133
Figure 4-3. Illustration of individual ancestry estimates along a chromosome .......................... 135

Figure 4-4. Relationship between skin pigmentation and Indigenous American ancestry in sample .................................................................................................................................................. 138

Figure 4-5. Collection locations of samples used in linkage analysis ........................................ 141

Figure 4-6. Distribution of Ancestry and Melanin ........................................................................ 144

Figure 4-7. Distribution of Genotypes for Selection Nominated Candidate Genes .................... 149

Figure 4-8. Regressions of melanin index (M) versus Indigenous American ancestry in four populations .................................................................................................................................................. 151

Figure 4-9. Distribution allelic association Z-scores across genome in Colombian and San Luis Valley populations .................................................................................................................................................. 154

Figure 4-10. Z-scores for allelic association on chromosome 15 .................................................. 155

Figure 4-11. Distributions of adjusted M index by genotype at genes of interest ....................... 164
LIST OF TABLES

Table 2-1. Summary of previous indications of selection at pigmentation genes.................35

Table 2-2. Sample populations and sizes included in meta-populations from HGDP sample. ......................................................................................................................43

Table 2-3. Percentages of SNPs/windows showing statistically significant results for each test in each population ........................................................................................................64

Table 2-4. Comparison of Selection in Pigmentation and Non-Pigmentation Genes .............86

Table 3-1. Mean differences and p-values for Welch’s t-tests.................................................113

Table 3-2. Comparison of ancestry estimates for each ethnic group based on autosomal and NRY markers........................................................................................................115

Table 4-1. P-values for Kruskal-Wallis Tests......................................................................159
ACKNOWLEDGEMENTS

First and foremost I would like to thank my parents – Marc Quillen and Marilyn Harp – who supported me from the very beginnings of my love for science, even through the phase when their land-locked daughter decided her dream was to be a marine biologist. They and my brother Kevin nurtured my intellectual curiosity and cheered me down every path I chose. Without them and the loving encouragement of Geoff Vasile, I would not have been able to complete this journey.

My advisor, Mark Shriver, has been a constant source of new ideas and has allowed and encouraged me to diversify my interests and pursue whatever questions I found interesting. His confidence in my abilities and willingness to allow me a great deal of independence in my research have given me the self-assurance I need to start my professional life. I have also received both thoughtful critique on my research and helpful guidance about my career from my committee – Nina Jablonski, David Puts, David Vandenburgh, and Ken Weiss. Along with Mark, these individuals have fostered an environment where I was encouraged to stretch and grow as a scientist.

My dissertation research could not have been completed without the support of a number of collaborators including Marc Bauchet, Carolina Bonilla, Tom Brushart, Bob Ferrell, Gerrardo Gutierrez, Meaghan Healy, Keith Hunley, Yann Klimentidis, Lorna Moore, and Mark Stoneking. Abby Bigham and Xianyun Mao also directly contributed their time and skills to help me understand and use the Mega data, saving me months of work. The research was funded by a National Science Foundation Doctoral Dissertation Improvement Grant (#0925976), Hill and Baker Grants from the Department of Anthropology, and a RGSO grant from the College of Liberal Arts at the Pennsylvania State University.
There are a number of friends who have been wonderful sources of intellectual and moral support in my lab – Laurel Pearson, Abby Bigham, Denise Liberton, Jen Wagner, and Kerri Matthes – in my department – Steve Beckerman, Jason DeLeon, Sharon DeWitte, Holly Dunsworth, Brenda Fraser, Kirk French, Craig Goralski, Logan Kistler, Jeff Kurland, Erick Rochette, and Sam Sholtis – in State College – María Inclán, Dawn Miller, and Sarah Rochette – and elsewhere – Veronica Davis, Summer Eglinski, LeighAnn Mullin, and Sarah Taylor. To adequately express my appreciation for each of these individual would take more pages than the dissertation itself so suffice it to say that all of these people have made my time at Penn State a wonderful period of my life, full of both fun and learning, for which I am forever grateful.
Chapter 1

Introduction and Background

The evolution of human pigmentation mirrors the history of the species. As Homo sapiens spread across the world, pigmentation was shaped rapidly by new environmental pressures, mutation, drift, and sexual selection. The results of these evolutionary forces can be seen in modern genomes where genes known or suspected to play a role in pigmentation are among the most rapidly evolving in European, East Asian, South Asian, African, and – as will be discussed – Indigenous American populations [1-7]. The speed of these changes and the extremes in phenotype seen across modern populations should make pigmentation among the easiest complex traits to fully understand, but the picture remains incomplete. This research addresses one of the major gaps in current knowledge by focusing on genes responsible for the mean differences in pigmentation between Indigenous American and European populations.

The genetic study of human pigmentation is fundamental to anthropology because pigmentation represents an enormously variable phenotype present in all humans, shaped and reshaped by the forces of evolution over the history of the hominid lineage. Although any two human chromosomes are, on average, more than 99% identical, among the 1% that varies are genes responsible for striking differences between individuals including a range of skin colors from porcelain to ebony, eyes that come in blue, green, brown and mixes of all three, and hair that varies from blonde to black to shocking orange. Pigmentation is among the most conspicuously varying traits among individuals and populations and must be understood in the context of the evolutionary history of the human species.

Contemporary variation in skin pigmentation indicates that skin has responded to environmental changes as human populations lost their fur, moved into the more northern climates of Europe and Asia, and then migrated into the Americas. The cycle of lightening and
darkening that appears to have happened in the evolutionary history of Indigenous American populations will be the focus of this study. Humans are unique among apes in occupying a broad environmental range and varying environmental pressures are the leading explanation for the phenotypic and, at some loci, genetic differences among populations within our species. As a population moves into a new environment, selection acts on both existing and newly emerging genetic variation in the population. As a result, convergent evolution among continental groups, different genes contributing to similar phenotypes, may be common for pigmentation genes. Convergent evolution of light skin pigmentation has already been identified in European and East Asian populations [8] and may also contribute to the evolution of dark skin pigmentation in equatorial regions around the world.

While variation in skin pigmentation across the globe can be directly observed and the spread of the human species out of Africa, across the Old World, and into the Americas is well documented in the paleoanthropological record and through genetic studies, the genetic mechanisms through which evolution altered this phenotype remain unclear. Approximately 76 potential candidate genes for variation in human pigmentation have been identified through clinical disorders, whole genome association studies, and orthologous genes in model organisms including zebra and stickleback fish, mice, and pigs, but only a few genes have been shown to play a substantial role in pigmentation differences among populations. These genes will be reviewed in this and the following chapters.

To narrow the list of candidate genes to those most likely to be responsible for differences among populations, a survey for signatures of selection on the candidate genes was performed. Examinations of genomic signatures of selection have previously been completed for lightening of skin pigmentation in European and East Asian populations [1, 2, 7, 9] but not yet for Indigenous American populations. Therefore, the first component of this research (Chapter 1) has been to remedy this lack of information on patterns of selection in pigmentation genes in
Indigenous Americans. To generate these preliminary results, three tests of selection – Locus-Specific Branch Length (LSBL) [10], Reduced Heterozygosity (LnRH) [11, 12], and Tajima’s D [13] – were used to identify pigmentation genes with signatures of selection unique to Indigenous American populations. While evidence of selection does not indicate that the genes have a functional role in pigmentation, this method can be used to prioritize a subset of genes for investigation. These will be referred to as “selection-nominated candidate genes” [9].

Based on the evidence of selection in these genes, a logical hypothesis is that they may play a role in determining skin color in Indigenous American populations. This hypothesis was tested in the second phase of the research where these selection nominated candidate genes were genotyped in admixed populations from the Americas. In an admixed population, genes influencing skin pigmentation variation between the Indigenous American and European parental populations will have a higher proportion of Indigenous American alleles than expected in individuals with darker skin pigmentation. Characterization of one of the samples included in this analysis, a group of Mestizo, Indigenous, and Afro-Colombian individuals collected in southwestern Colombia, using autosomal and Y chromosomal ancestry informative markers (AIMs) will be discussed in Chapter 3.

By using admixed populations we also reduce the likelihood of false-positive associations between skin pigmentation and any loci showing allele frequency differences by conditioning the associations on locus ancestry. Five-hundred and fifteen individuals with mixed Indigenous American and European ancestry were genotyped for single nucleotide polymorphisms (SNPs) chosen to be ancestry informative between the European and Indigenous American parental populations. These SNPs were used to test for admixture linkage association between the candidate genes and variation in skin color. These results are discussed in Chapter 4.
Importance of the Study of Human Skin Pigmentation

Pigmentation has been a subject of formal anthropological study for more than a century and humans have been documenting the great variety of pigmentations present in our species for over three millennia. The tomb of the Egyptian pharaoh Seti I shows four men, thought to be fair-skinned Asian and Libyan men, an extremely dark-skinned Nubian, and an intermediately pigmented Egyptian [14]. A fuller understanding of the evolutionary genetics of pigmentation will contribute broadly to the understanding of the evolutionary history of our species. Specifically, this research will enhance the study of pigmentation genetics in understudied New World populations.

Additionally, the study of skin pigmentation is of interest as a model for the interaction between genes and a fluctuating environment. Convergent evolution has already been shown to produce lighter skin between humans and zebrafish due to homologous genes and between Europeans and East Asians due to mutations in different genes [8, 15]. Asian populations, probably bearing mutations for lighter skin, carried their evolutionary history with them into the New World as a sort of genetic legacy. As their descendents migrated into Mesoamerica, Central America, and the northern regions of South America, they encountered new climates with high levels of ultraviolet radiation which favor darker skin pigmentation. As a result of this expansion, the genetic make-up of these populations certainly evolved in many ways including in genes related to skin pigmentation. Alternately, some researchers argue that the populations migrating into the Americas could have retained the darker skin of their African ancestors and compensated for the decreased vitamin D production by consuming a high-vitamin D diet [16]. These hypotheses can be tested once the mutations involved in the lighter pigmentation of East Asian populations and the darker pigmentation of Indigenous American are better characterized. Regardless of the phenotypes of the populations entering the Americas, this study contributes to a
better understanding of the genetic changes caused by the new environments that these populations experienced.

A clear understanding of the genetic variation and differing evolutionary histories that have produced visible differences, such as skin pigmentation, between human populations not only benefits anthropologists and geneticists but also serves to combat the folk concept of “race” that aligns pigmentation and other physical traits with behavioral and cognitive traits and abilities. An evolutionary perspective that emphasizes that humans are all one species shaped by evolution in response to varying environmental pressures, in addition to stochastic events like mutation and genetic drift, can lead to a better understanding of the falsity of racial thinking. Because the rate of evolution differs between genes, a rapidly evolving trait like skin pigmentation is an ideal example of the way that selection has and will continue to act on the variation present in our species. The tangible nature of a superficial trait like skin pigmentation makes it an accessible learning tool for students. Additionally, the study of rapidly evolving genes is of interest to scientists as a contrast to the majority of genes which have experienced much less change in allele frequencies over the same time scale and in the same populations.

Lastly, an understanding of the genetic basis of physical traits is of interest to forensic investigators who could use the resulting data to predict the skin color of perpetrators who leave DNA at the crime scene or victims who have left only skeletal remains. The growing subfield of molecular photofitting uses genetic evidence to generate a likely physical profile of an unidentified individual. Previous work has focused largely on the use of genomic ancestry as a proxy for a wide variety of physical traits including skin, hair, and iris color as well as facial features, but as the genetic architecture of these traits is elucidated, the use of this information in molecular photofitting will become more common allowing for more detailed and individually specific predictions of phenotypes. [17]
Worldwide Distribution of Human Skin Pigmentation

The history of the spread of humans throughout the world and the manner in which populations adapted to the new environments they encountered is a primary subject of anthropological interest. The anthropological study of variation in skin pigmentation is as old as biological anthropology, but the first attempts at quantifying skin pigmentation came in the late 19th century. VonLuschan tiles were developed in 1897 to aid in the documentation of skin pigmentation around the world by comparison of skin to a set of 36 portable glass tiles [18]. In the 1950s, reflectance-based measurements became the standard, allowing for objective, quantitative measures of skin pigmentation [19]. Hundreds of thousands of volunteers have since had their skin pigmentation measured by eager biological anthropology graduate students (and other researchers) to produce a picture of the distribution of skin pigmentation throughout much of the world. This picture reveals that the degree of variation in skin pigmentation is apportioned approximately 88% among regional groups as opposed to local variability [20], a finding in stark contrast to the approximately 85% of overall genetic variability that is found within, rather than among populations [21, 22]. The unusual pattern of variation in skin pigmentation suggests that this phenotype, unlike the majority of human variation, may have undergone strong directional selection unique to distinct populations in addition to genetic drift as human populations spread throughout the world.

As the volume of data increased, a number of authors noted that there was a general relationship between increasing distance from the equator and lighter skin pigmentation [23, 24], including geographer Renato Biasutti who, in 1959, published a map of skin pigmentation around the world showing that more darkly pigmented populations live in the tropics and lighter skinned populations live closer to the poles (Figure 1-1).
The majority of researchers argued that this could be best explained by the decreasing intensity of ultraviolet radiation (UVR) with decreasing latitude [26-29]. In 1997, Relethford tested the relationship between skin pigmentation and latitude and found a strong correlation between skin pigmentation and latitude in Old World populations [30]. However, the slopes of these correlations are different in the Northern and Southern hemispheres as well as between men and women. Relethford contends that the hemispheric differences were due to more intense ultraviolet radiation in the Southern hemisphere although Chaplin and Jablonski noted this effect could have to do with the distribution of land masses, and therefore humans, in the two hemispheres [31]. In 2000, Jablonski and Chaplin supported this argument by using satellite measurements of ground-level UVR intensity to show that the relationship between UVR and skin pigmentation is stronger than the relationship between skin pigmentation and latitude alone [32]. Figure 1-2 shows the general pattern of UVR on the earth’s surface as well as predicted and observed world-wide skin pigmentation. The most recent work by Jablonski and Chaplin on the global distributions of ultraviolet radiation [33], emphasizes the different distributions of UV-A
(400-315 nm) and UV-B (315-280 nm). UV-B has a weak effect outside of the tropics and more seasonal variation in intensity. UV-A is also weaker outside of the tropics but is relatively more uniform across the year. The differences between these distributions are important because UV-A penetrates more deeply into the skin and the two forms of UVR have distinct impacts on biological functions which differentially create potential selective forces which may act to favor darker skin pigmentation where UV-A is highest and lighter skin pigmentation where UV-B is lowest.
Figure 1-2. Distribution of ultraviolet radiation and skin pigmentation world-wide. (a) General pattern of annual average ultraviolet radiation (UVR) on earth’s surface follows expectation that equatorial regions should have the highest levels of UVR. However, ground level factors alter the observed levels. High altitude regions, including the Andes, Tibetan Plateau, and East African Plateau have more UVR than expected for latitude while the Amazon rainforest decreases the ground-level UVR. (b) Predicted levels of skin pigmentation based on UVR levels. (c) A more accurate map of observed skin pigmentation with arrows indicating locations of sample populations. Figures from [32].
Hypotheses for Selective Pressures on Skin Pigmentation

Ultraviolet radiation interacts with skin pigmentation in a number of ways that have the potential to influence evolutionary fitness. In all cases, individuals with dark skin (high melanin levels) in high UVR environments and light skin in low UVR environments are at an advantage. A number of hypotheses have been put forth to explain the variation of skin pigmentation with latitude. Many of the earlier hypotheses, some based on conjecture, others on data that proved later to be misleading, have been discredited. [34] Among these are the idea that darker skin pigmentation evolved as camouflage and that lighter skin pigmentation afforded protection from frostbites [35, 36]. A number of other hypotheses have remained and those with substantial evidence to support them are reviewed here.

Ultraviolet Radiation and Skin Cancer

The dangers of sunburn and melanomas are one pressure which could link increased UVR to darker skin pigmentation. For many people, a sunburn is merely an annoyance, but repeated sunburns, especially at a young age, greatly increase the likelihood of melanomas, a particularly rapidly metastasizing and lethal form of cancer which could shorten the reproductive lifespan of an individual [37]. Several studies of individuals living in equatorial climates have found precancerous growths in more than 90% of those with albinism compared to almost no cases of such growths in darkly pigmented individuals [38]. In modern times, people usually have the opportunity to spend at least some part of the day indoors, a means of managing UVR damage that early hominids would not have had. Melanin, located in the basal layer of the epidermis, acts as a photo-filter which scatters UV-A so that it cannot damage the DNA of the lower dermis, protecting the individual from skin cancers. Additionally, sunburned skin may
become infected, decrease range of motion, and limit the skin’s ability to sweat which is the primary means of thermoregulation in the human body. [17] Limited range of motion would also decrease the ability to perform life-sustaining tasks like breast-feeding, hunting and gathering, and escaping from predators. Several genes known to contribute to variation in skin have also been linked to an increased risk of skin cancers – both melanomas and basal cell carcinomas. [39] However, this link does not imply that these genes were under selection in response to skin cancer risk. Particularly because the majority of the polymorphisms investigated show evidence of skin lightening associated with populations moving into Europe and not with the much earlier skin darkening that would have protected our earliest ancestors from equatorial UV-A. While melanomas and sunburns are a risk for lightly pigmented modern humans, it is unclear to what degree the detrimental effect of melanoma would shorten the reproductive lifespan of early hominids. [32]

**Ultraviolet Radiation and Folic Acid Photolysis**

There is also strong evidence for a selective advantage for dark skin to avoid folic acid photolysis in high UVR environments. Degradation of folate in vivo in light-skinned humans has been documented following exposure to natural sunlight [40] and specifically to UV-A. Dark epidermis protects against folate photolysis, which reduces the risk of neural tube defects (NTDs) such as spina bifida [41]. Without medical intervention, NTDs would virtually always be lethal due to nerve damage, paralysis, and infection. NTDs are more common in light-skinned populations where neural tube defects accounted for up to 15% of all perinatal and 10% of all postperinatal deaths prior to the introduction of pre-natal supplements [42].

Despite observations of a relatively high rate of infant death due to neural tube defects in light-skinned populations, some researchers contend the rate of neural tube defects would have
been too low to have a substantial influence on fitness. Although NTDs are likely the most lethal disorder related to folate photolysis, an increasing number of other health problems are being linked to folic acid deficiencies. Work in model organisms suggests that a lack of folate may inhibit spermatogenesis resulting in male infertility [43, 44]. Folate deficiencies have also been linked to heart disease and cancers [45]. Folate is also essential in proper DNA replication and repair, a particular concern in high UVR environments where UV can cause pyrimidine dimers in nuclear DNA of the skin [46-48].

**Alternate Theories Favoring Dark Skin Pigmentation**

Elias et al. [49] contend that in addition to the UVR damage, a major risk following hominid loss of body hair would have been xeric stress – or a loss of moisture in the skin in an increasingly arid environment. One of the primary functions of skin is undoubtedly as a barrier between the internal homeostasis and external environment and the strength of this barrier lies in the stratum. Elias argues individuals with more melanosomes (causing darker skin pigmentation) have a stronger and more acidic stratum corneum. Under this hypothesis, darker skin pigmentation would have been a byproduct of selection favoring more efficient barrier functionality in the skin.

At this time, evidence for hypotheses other than intense UVR causing the darkening of skin pigmentation in early hominids is weak. Whatever the cause, the evidence for selection in favor of darker skin closer to the equator is strong. In addition to the studies indicating adaptive benefits for increased melanin in high-UVR regions, molecular evidence supports constraint on variation in skin pigmentation early in our lineage. Although it can be difficult to identify signatures of selection around the time early humans lost their fur, molecular evidence points to
purifying selection removing mutations from the pigmentation gene *MCIR* in equatorial African populations since our hominid lineage split from chimpanzees [50].

**Ultraviolet Radiation and Vitamin D Production**

In far northern and southern climates, decreased UVR intensity relaxes the selective pressure in favor of darker skin but this relaxed selective environment alone is likely insufficient to account for the rapid and apparently independent change from darker to lighter skin. Genetic studies have shown that lightening of skin has occurred independently in European and East Asian populations [8]. The fixation of alternate functional alleles in the gene *SLC24A5* in European and East Asian populations since the two populations split suggests a relatively rapid increase in frequency. Skin with high concentrations of melanin becomes maladaptive at higher latitudes because UV-B radiation is a necessary component in the production of vitamin D₃ which is essential for normal growth, calcium absorption, and skeletal development [51]. On average, individuals with greater constitutive skin pigmentation have lower levels of circulating vitamin D than lighter-skinned individuals living in the same environment [52]. A deficiency in vitamin D₃, called osteomalacia in adults and rickets in children, can cause pelvic deformations in women that prevent normal childbirth [53] and recent dark-skinned immigrants to higher latitudes are at a higher risk than natives of vitamin D₃ deficiency [54]. Children with severe rickets are frequently bow-legged and have difficulties walking. Vitamin D has also been implicated in a variety of disorders involving the improper functioning of the immune system, cancer, cardiovascular disease, and potentially others [55-57]. More than sixty tissues in the human body have vitamin D receptors suggesting that there may be many factors influencing the ideal levels of vitamin D production. Each of these pathways potentially contributes to the selective pressure in favor of the maintenance of vitamin D production in the skin. [58] Most importantly, vitamin D plays a
role in the prevention of disorders that influence not only long-term survival but fertility and early childhood development which are critical to evolutionary fitness [59].

This hypothesis that lower levels of constitutive melanin evolved in high latitude populations in response to a decreased ability to produce vitamin D has been challenged by a number of researchers who have cited the availability of vitamin D from dietary sources, the unclear association between cutaneous melanin and serum vitamin D levels, and the severity of health effects due to vitamin D deficiency. Vitamin D$_3$ can be found in some animal foods such as oily fish and eggs while vitamin D$_2$ can be obtained through some plant-derived sources [16]. However, in worldwide human populations, more than 90% of vitamin D is produced by cutaneous photosynthesis [60]. A meta-analysis of studies on the relationship between melanin and vitamin D production following exposure to UV-B found that studies where only one exposure was done found increased melanin decreased vitamin D production where those with multiple UV-B exposures found no association [61]. Furthermore, Bogh et al. [62] found that 85% of healthy Danes were either deficient or insufficient in serum vitamin D levels in the winter, suggesting that seasonal vitamin D deficiency may not lead to adverse health effects. Additionally, they concluded that the increased vitamin D response following UV-B exposure was more closely correlated with serum cholesterol levels (which share a similar pathway to vitamin D production) than constitutive skin color. Determining the relationship between vitamin D and skin pigmentation is additionally complicated by variation in vitamin D receptors which are encoded, in part, by the genes $GC$ and $VDR$ which show substantial polymorphism and vary across populations [63]. Recently, Robins [60] revived an earlier argument [27] against the primacy of rickets as a primary evolutionary force driving lighter skin pigmentation on the basis of studies which have found that certain groups (frequently of Asian ancestry) are at higher risks of rickets than more darkly pigmented groups (usually of African ancestry) living in the same region.
In response, Chaplin and Jablonski [64] noted that the correlation between skin color and UV-B exposure has an $R^2$ value of 0.927, that insufficient vitamin D levels from both diet and lack of solar exposure are still causing rickets in some parts of the world today, and that even if rickets is not the primary selective pressure favoring increased vitamin D production there is still a rapidly expanding literature on other medical costs to low levels of vitamin D. Although the question of the primary cause of skin lightening in humans at higher latitudes may not be fully resolved, as with the skin darkening, the evidence at the level of the genome is compelling. As discussed earlier, convergent evolution of lighter skin pigmentation in European and East Asian populations, and even with our Neandertal relatives [65], has been clearly documented.

**Sexual Selection in Skin Pigmentation**

A final factor when considering the evolutionary history of human pigmentation, including skin as well as hair and iris pigmentation, is the role that sexual selection may have played in altering these highly visible traits. Sexual selection could favor either darkly or lightly pigmented individuals in any type of environment. Darwin was the first to propose a role for sexual selection in shaping human pigmentation in *The Descent of Man and Selection in Relation to Sex* [66]. The ability for sexual selection to cause substantial changes in populations, even to the point of speciation has been modeled [67], but the ability to detect sexual selection – or, more accurately, to differentiate between natural and sexual selection – based on statistical analyses has eluded scientists. Sexual selection is frequently invoked to explain the variation seen among populations at the same latitude or among closely related populations where natural selection cannot explain the variation [68]. However, as with studies of natural selection, differentiating this variation from drift is difficult as many of the populations studied are small.
Aoki [69], among others, contends that sexual selection is most likely to contribute to depigmentation at latitudes where the selective pressure favoring darker skin pigmentation is relaxed. As evidence, Aoki points to the large number of variants related to lighter skin pigmentation in the MC1R gene found in Europe as compared to the relative lack of variation in Africa. This is counter to the expectation of more genetic diversity occurring in Africa and supports a hypothesis of purifying selection at this locus in dark-skinned populations. Intrinsic to a claim of sexual selection must be a preference for lighter skinned mating partners. Van den Berghe and Frost [70] reviewed evidence from the Human Relations Area Files and found that in 47 of 51 populations where a preference was reported, lighter skin pigmentation was preferred. This preference is even stronger among males than among females. Based on this evidence, Frost later argued [71] that variation in hair and eye color, and possibly depigmentation, in Northern European women could be due to sexual selection in hunter-gatherer populations where males were rare. The intrinsic difficulty in associating any previous changes in allele frequencies with particular causes is that all of the evidence is circumstantial and based in correlations. Sexual selection is even more difficult to document than natural selection because it is based on preferences that can change even faster and with less of a trace than environmental pressures.

One line of evidence indicating a potential role for sexual selection in shaping human skin pigmentation is the mean pigmentation difference seen in many populations between males and females. For example, a study of six islands in Melanesia found that males had significantly darker skin than females on five of the islands [68]. Similarly, Relethford [30] identified different correlations between latitude and skin pigmentation between males and females with female skin becoming lighter more rapidly with increasing distance from the equator. However, this may be due to the higher demands for calcium in females than males and the associate need for more vitamin D production [34]. A meta-analysis of human sexual dimorphism in skin pigmentation examined 53 cases in which skin pigmentation levels for males and females of the
same age were reported but failed to find significant differences and could not support the sexual selection hypothesis [72].

**Genes Involved in Melanogenesis**

There are many genes potentially involved in the vast variation seen in human pigmentation, many of which could be subject to selection in various populations. Skin color is predominantly determined by the pigment melanin, although other pigments, primarily carotene, reduced hemoglobin, and oxyhemoglobin, can also influence it. Melanoblasts in the dermis differentiate into melanocytes which are located on the basement membrane at the epidermal-dermal junction where they synthesize melanin. Melanocytes are closely associated with keratinocytes, to which they transfer melanosomes, the lysosomally-derived cytoplasmic organelles where melanin synthesis occurs. [73] The amount and type of melanin synthesized in the melanosomes and the number, size, and distribution of melanosomes are the predominant determinants of pigmentation [74]. *OCA2* and *SLC24A5* in particular influence melanosome number and proliferation leading to phenotypic variation in pigmentation [15, 75-77]. Figure 2-3 shows the variation in melanosome distribution in three populations.
Figure 1-3. Melanosome distribution in three populations. In lightly pigmented individuals like the Europeans and East Asians shown here, the melanosomes are less pigmented, smaller, and packaged into groups. In darker skin, melanosomes contain more melanin, are larger, and are distributed singly instead of in groups. Figure from [78].

Every step in melanogeneis is controlled by genes with the potential to affect the ultimate phenotype. These genes include transcription factors, membrane and structural proteins, enzymes, several kinds of receptors and their ligands, and they control a variety of functions including melanoblast proliferation, migration and survival, melanosome function, and melanogenesis. A number of genes involved in the production of melanin are also involved in other pathways, particularly during development and in the immune response. Although these alternate functions could place selective constraints on the genes in question, this discussion will be constrained to pigmentation pathways. Figure 2-4 illustrates the roles of many of the genes important in melanin production.
There are two main forms of melanin which differ biochemically and structurally - eumelanin, a brown/black pigment, and pheomelanin, a yellow/red compound. The initial biochemical step in the synthesis of both forms of melanin is the oxidation of tyrosine into dopa, which is further oxidized to dopaquinone. Both reactions are catalyzed by the enzyme tyrosinase, the gene for which (TYR) shows substantial allelic variation between populations [79]. MATP has been implicated in the movement and processing of tyrosinase [80]. From dopaquinone on, biosynthesis of eumelanin and pheomelanin follow different routes. MC1R, one of the first genes
found to cause variation within European skin and hair color [81], and ASIP [82] regulate the relative production of eumelanin and pheomelanin respectively. In total, 76 genes have been identified as candidate genes for variation in human pigmentation.

In addition to what is now known about the molecular basis of pigmentation, researchers have found that a number of the genes in the melanin pathway are variable among populations. Studies of skin color in individuals of known degrees of admixture indicate that ten or more genes may be responsible for the variation in skin pigmentation between West African and West European populations [83]. At present, SLC24A5, MATP, KITLG, OCA2, TYR, and ASIP have been implicated in this variation using admixture mapping [15]. Based on admixture linkage to an ancestry informative marker in CYP19A1, MYO5A and SLC24A5 have been proposed to cause skin color differences between Indigenous Americans and Europeans [84]. These functional and admixture based studies have been supplemented in the last several years by several studies of the genomic evidence for selection at pigmentation genes which are reviewed in the following chapter.

**Skin Pigmentation in the New World**

The current body of knowledge regarding skin pigmentation in the Americas has many gaps which have been filled with a several opposing hypotheses about the skin color of the earliest Americans and what subsequent changes may have occurred in these populations. When humans entered the New World, the environment changed from the low UVR found in northern Asia to higher levels of UVR in Central America. The fluctuating evolutionary landscape through which humans passed before occupying the Americas has likely left signposts in the genomes of modern Indigenous Americans in the form of evidence of selection of varying ages
and intensities. This evolutionary legacy makes the evolution of skin pigmentation genes in Indigenous Americans particularly interesting to study.

The likely skin color of the first Americans is somewhat controversial for all of the reasons involved in the controversy over the strength of selection conferred by the need for cutaneously-produced vitamin D. Logically, if low levels of UV-B select for lighter skin pigmentation, the ancestors of the first Americans must have had light skin to survive in Central and Northern Asia. However, it has been argued that a diet like that found among the Inuit and Evinki (high in caribou and reindeer that eat lichen, and small oily fish) would have eliminated the selective pressure in favor of lighter skin pigmentation at a high latitude. The amount of vitamin D in many fish and other seafood is sufficient to compensate for the body’s inability to produce it without extreme amounts of food needing to be consumed [16]. Proponents of the hypothesis that the first Americans were already darkly pigmented point to the relatively dark skin pigmentation found in Inuit populations [51]. Contrary to this evidence is the gradient in skin pigmentation found in the Americas (see Figure 1-1) where the majority of populations living closer to the poles are lightly pigmented and individuals closer to the equator are more darkly pigmented. Modern Indigenous American populations from high UVR regions are, on average, more darkly pigmented than both East Asian and European populations. Unpublished samples from this and other research show that U.S. residents of East Asian ancestry sampled at Pennsylvania State University have a mean melanin index of approximately 34 ± 4.5, European Americans sampled at the same location have a mean of 30.5 ± 3.0, and a sample of Indigenous Americans from Guerrero, Mexico have a mean of 46.1 ± 4.9.

Though the phenotype differences have been documented [85], no genes involved in these differences had been identified prior to this research. The identification of alleles favored by selection that contribute to darker skin pigmentation in equatorial Indigenous American-
European admixed populations would bolster the evidence in favor of the evolution of darker skin pigmentation in modern Indigenous Americans from a lighter ancestor.

The vast body of research on skin pigmentation in human evolution has led to a detailed, though incomplete, understanding of this phenotype. Many key issues remain and this research addresses two of these questions. Chapter 2 discusses the ability to detect selection that may have occurred over in Indigenous Americans since their split from their common ancestors with East Asian populations. By using the same statistical methods on different populations within the same regions which were genotyped using different genome-wide panels of markers, this study also addresses the robustness of these methods and the substantially different results that may be obtained due to chance differences in sampling and marker selection. In Chapter 4, the genes identified in Chapter 2 are tested in admixed populations to determine the role that these genes play in skin pigmentation differences between Indigenous American and European populations. Previous researchers have incompletely addressed the important issue of whether or not identifying signatures of selection is equivalent to identifying functionally important genes contributing to differences among populations. The study of skin pigmentation in New World populations is an important first step in completing our understanding of human variation in skin pigmentation.
References

Chapter 2

Evidence for Natural Selection at Pigmentation Genes in the Americas

Among the most important goals of anthropological genetics is to identify the genes responsible for major phenotypic changes in human evolutionary history. One means of doing this is to locate genes that have recently undergone selection in different populations. The modern distribution of skin pigmentation has certainly been influenced by the decreased variability and drift resulting from the repeated genetic bottlenecks that characterize our species’ demographic history. However, the evidence reviewed in the previous chapter indicates that natural and sexual selection have also been important. A model based on UVR as a selective force combined with what is known about the evolutionary history of the human species produces a series of testable hypotheses about changes in skin pigmentation as humans adapted to new environments. With the exception of the face, lips, pinnae, palms, and anogenital areas, the majority of primates have light or completely depigmented skin under their hair [1-3]. Parsimony suggests the primitive state of skin pigmentation in the hominid lineage was light skin and as our ancestors lost their fur, they likely evolved darker skin in response to the effects of UVR on the newly bare skin. As H. sapiens migrated out of Africa, they were probably dark skinned in response to living in a high-UVR environment and some common changes [4] occurred that are now shared by Europeans and East Asians. However, once humans populated Europe and Asia, independent genetic changes resulted in lighter skin pigment in both of these populations [5]. Interestingly, the preliminary sequencing of the Neandertal genome [6] led to the identification of independent mutations in the MC1R gene, not found in humans today, which would have similarly lightened skin and hair pigmentation among our evolutionary cousins [7].

Because skin pigmentation has undergone several stages of darkening and lightening as populations spread throughout the world, evolutionary forces have acted in different ways in
different populations resulting in both population-specific and shared allelic variation at many pigmentation genes. Geneticists have developed a number of statistical tests to detect evidence of natural selection at pigmentation genes and applied these tests to a several Old World populations. The results of this research are summarized below.

**Methods for Identifying Selection in the Genome**

The degree to which skin pigmentation is characterized on the genetic level establishes a clear set of candidate genes for analysis of selection in different populations. Over the past decade, as large scale data sets of human populations have become more massive, less expensive, and more widely available, many researchers have applied a variety of different tests of selection to this new genome-wide data. By characterizing the evidence for selection in geographically dispersed populations, the well-documented history of the human expansion throughout the globe can be married to the genetic changes that resulted from new selective pressures in the environments encountered during human expansion. In theory, the ability to detect selection in a population can persist for a number of generations equal to the effective population size of the population [8]. This means that although a population may have a large census number today, the ability to detect ancient selection may be limited due to previous bottlenecks.

When considering selection in the genome, we generally include three types of deviations from neutral expectations – directional (positive), balancing, and background (negative or purifying) selection. Each of these types of selection has a different effect on the genome. In the cases of drift under neutrality, directional selection, and background selection, a single allele may increase in frequency or reach fixation. However, these forces can be distinguished in the genome by the speed with which a particular allele reaches fixation which can be inferred from the amount of variation surrounding the gene under selection. Figure 2-1 illustrates the how these
unique forces influence gene trees. For the purposes of this research, only directional selection will be considered as it is expected to be the predominant type of selection on pigmentation phenotypes in a population entering a new environment. However, it is also important when discussing skin pigmentation to consider not only the onset of selection but also the removal of purifying selection which may have been occurring in the previous environment. The ending or diminishing of purifying selection may allow one or more alleles to increase in frequency due to drift in such a way that they appear to be unique to the new population.

Figure 2-1. Effects of natural selection on gene trees. (a) Even under a neutral model, an allele can drift to fixation. This is particularly common when effective population size is small as in genetic bottlenecks like the one that occurred when humans colonized the Americas. (b) Positive selection can be distinguished from neutrality because fixation occurs much more rapidly. In addition to the large allele frequency differences that will occur, selection will decrease the variation in the regions surrounding the gene. Methods for identifying positive selection are discussed in the text. (c) Balancing selection will maintain multiple alleles in the population over a long span of time and the alleles will have a deep coalescence. This deep coalescent time means that there will likely be more variation surrounding alleles under ongoing balancing selection than positive selection. (d) Because background selection does not favor a single allele but instead removes deleterious alleles from a population, it can be distinguished from the other forms of selection. Figure from [9].

The identification of selection in the genome is predicated on the ability to detect patterns in the genome caused by selection not by other forces like genetic drift, gene flow, or mutation.
Separating these evolutionary forces is the greatest challenge that must be overcome to successfully identify selection in the genome. This issue is especially important when considering Indigenous American populations which are known to have had small effective population sizes and therefore experienced more drift [10].

Selection has locally specific effects on the functional loci, while the other forces are expected to be uniform across the genome. There are many methods of identifying selection in the genome, but they all center around a few main concepts. When selection favors a particular phenotype, it will drive an increase in functional mutations in genes related to the phenotype as compared to genes not undergoing directional selection. Because amino-acid altering (non-synonymous) mutations are usually deleterious, background selection will usually eliminate them from the genome. An excess of such non-synonymous changes can be identified using techniques such as the McDonald-Kreitman test [11]. This test has a number of limitations including only detecting changes in the coding regions of genes and not in any regulatory regions that could influence gene expression. Because this test requires multiple selected non-synonymous changes to clearly differentiate the selected gene from those evolving neutrally, this test is not ideal for comparisons between recently diverged populations like the Indigenous Americans and East Asians.

Directional selection also alters the genome by removing variation in the gene under selection and at linked loci surrounding the gene. Tests such as Tajima’s D [12], the Hudson-Kreitman-Aguadé (HKA) [13], Fu and Li’s D [14], and Fay and Wu’s H [15] are all based on identifying this reduction in genetic diversity and the resulting large number of rare alleles that will occur as new mutations arise following the so-called “selective sweep.” Tests such as CLR (composite likelihood ratio) [16] consider both the local reduction and substantial allele frequency differences. When comparing between populations, an excess of high-frequency derived alleles, sometimes referred to as a skewed site-frequency spectrum, may indicate
selection. It can also be useful to simply examine the large allele frequency differences that result when selection acts on one of two reproductively isolated populations. Tests like $F_{st}$ [17], FR ($F_{st}$ ratios) [18], and Locus-Specific Branch Length (LSBL) [19] can be used to identify the extensive population differentiation in genes that may be undergoing directional selection.

These extreme differences in allele frequencies will also reduce the amount of variation at individual loci within a population under selection so that tests like the log of the ratio of variance ($\ln RV$) [20] and ratio of heterozygosities ($\ln RH$) [21] can be used to detect this reduced variation when the object populations is compared to a reference population. Most recently, classes of tests have been developed that focus on identifying long haplotypes surrounding genes that have undergone selection. These include Extended Haplotype Homozygosity (EHH) [22] and its derivatives SNP-specific EHH (EHHS) [23] and cross-population EHH (XP-EHH) [24] as well as integrated haplotype score (iHS) [25] and haplotype block length ratio (LR) [18]. This class of tests is particularly useful for identifying partial sweeps, where rather than going to fixation the frequencies of the alleles under selection may be as low as 10% [26].

The different methods of identifying selection have different time-depths over which they can detect events. This is an important consideration when the focus of the work is on population-level differences. These differences make some statistics more appropriate for considering Indigenous-American specific changes – which would have occurred in the last 20,000 years. As illustrated in Figure 2-2, the three tests conducted in this study – tests for rare alleles, reduced heterozygosity, and population differentiation – are all capable of detecting selection in the relatively recent time span during which Indigenous American populations diverged from Old World populations. The three tests used in this analysis are discussed in greater detail below.
When considering any statistical tests, it is important to consider the limitations and assumptions that underlie the methodology. The majority of these tests are designed to identify so-called “hard sweeps” where an adaptive mutation arises and immediately increases in frequency until it reaches fixation in a single population. This model is also applicable in cases where polymorphism exists at a very low level before the population enters a new environment where the rare allele becomes advantageous. [27] This could be the case for skin pigmentation if, for example, an allele for darker skin pigmentation is deleterious in northern Asian populations (and therefore selected against) but becomes advantageous when members of that population reach high-UVR regions of the Americas. Although a number of other pigmentation genes (most notable SLC24A5, KITLG and MATP) show evidence of hard sweeps, there is no way of guaranteeing that pigmentation alleles favored in the Americas would follow the hard sweep
trajectory. However, as the results below indicate, there are pigmentation genes that are outliers in the distributions for these tests of selection. Yet, it should also be acknowledged that there may be genes that play a significant role in inter-population differences in skin pigmentation that will not be detected by these types of analyses because they are not subject to a hard sweep.

In contrast to hard sweeps, “soft sweeps” are cases where selection on individual alleles may be relatively mild due to weak or temporally variable selection pressures or due to polygenic adaptation where the phenotype may be under strong selection but the selective pressure is distributed over many loci. The term soft sweep is also used to indicate a case where selection acts on existing variation rather than a newly arisen allele. When Coop et al. [28], found a paucity of high-$F_{ST}$ signals associated with long haplotypes in geographically dispersed populations, it was argued that hard sweeps – defined by Coop et al. as a new mutation sweeping to fixation with a selection coefficient of more than 1% – are rare in human evolution. Since skin pigmentation is known to be polygenic trait with gradual changes in selective pressure across continents, these arguments raise a concern. In a model with a large number of genes contributing to a phenotype, the selective pressures will only continue until the population’s phenotypic mean reaches the optimum which may not result in large shifts in individual allele frequencies. [27] There are many reasons to be cautious in using these tests of selection and the results should always be validated by experimental evidence as discussed in Chapter 3. However, there is substantial precedent for finding evidence of selection in skin pigmentation genes in human populations.

Selection at Skin Pigmentation Genes in Old World Populations

The results of the dozens of scans for selection across the human genome can be combined to represent a testable evolutionary-genetic framework for skin pigmentation that
mirrors the order of the separation of the major continental groups of the Old World. The combined results of twelve such studies can be found in Table 2-1. As can be seen from the table, evidence for selection is identified more frequently in non-African populations. This may be associated with the relatively recent colonization of new environments by some of these populations. Expansion into new environments usually involves a population bottleneck which decreases variation and increases linkage disequilibrium, making it easier to identify signatures of selection. However, increasingly evidence supports selection at skin pigmentation genes in equatorial populations in Africa which is to be expected considering the range of selective environments across the continent. Additionally, it is interesting to examine the pattern of selection as humans expanded out of Africa. A few genes show selection in both East Asian and European populations (e.g. KITLG) but there is also evidence of convergent evolution of light skin pigmentation in European and East Asian populations – meaning the shared phenotype does not result from a shared ancestry [5, 29]. It is important to note that no genome-wide study prior to this one has previously looked for evidence of selection at skin pigmentation genes in New World populations.
Table 2-1. Summary of previous indications of selection at pigmentation genes. To best illustrate overlapping results, populations have been condensed into simplified categories based on broad categories of biogeographic ancestry. To clarify, studies [18], [30], [31], and [32] utilized the Perlegen populations [33] which are more accurately described as European-American, African-American, and Chinese-American. [24], [25], and [34] use the HapMap populations as used in this research. [23], [29], and [35] used a combination of the HapMap and Perlegen or HGDP populations. [36] uses the HGDP samples described in this study and [37] uses set of individuals described as African-American, European-American, and Chinese.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Population</th>
<th>Studies</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM17</td>
<td>East Asian</td>
<td>[25, 30, 34]</td>
<td>Fst, EHH, LSBL, Tajima’s D, lnRH, iHS</td>
</tr>
<tr>
<td>ADAMTS2</td>
<td>East Asian</td>
<td>[25, 30, 34]</td>
<td>Fst, EHH, LSBL, Tajima’s D, lnRH</td>
</tr>
<tr>
<td>ATP7B</td>
<td>European</td>
<td>[35]</td>
<td>EHHS</td>
</tr>
<tr>
<td>ATRN</td>
<td>East Asian</td>
<td>[18, 31, 34]</td>
<td>LSBL, Tajima’s D, lnRH, CLR, FR</td>
</tr>
<tr>
<td>BNC2</td>
<td>East Asian, European</td>
<td>[34, 35]</td>
<td>LSBL, Tajima’s D, lnRH, EHHS</td>
</tr>
<tr>
<td>CYP2C8</td>
<td>East Asian</td>
<td>[35]</td>
<td>EHHS</td>
</tr>
<tr>
<td>DCT</td>
<td>African, East Asian</td>
<td>[18, 29, 30, 32, 34]</td>
<td>Fst, EHH, LSBL, LR, FR</td>
</tr>
<tr>
<td>DRD2</td>
<td>African, East Asian</td>
<td>[32]</td>
<td>EHH</td>
</tr>
<tr>
<td>DTNB1</td>
<td>European</td>
<td>[25]</td>
<td>iHS</td>
</tr>
<tr>
<td>EDA</td>
<td>East Asian, European</td>
<td>[25, 34]</td>
<td>LSBL, Tajima’s D, lnRH, Fst</td>
</tr>
<tr>
<td>EDN3</td>
<td>All Humans</td>
<td>[34]</td>
<td>Tajima’s D</td>
</tr>
<tr>
<td>EGFR</td>
<td>African, East Asian</td>
<td>[18, 32, 37]</td>
<td>EHH, LR, FR, Fst</td>
</tr>
<tr>
<td>HPS6</td>
<td>East Asian</td>
<td>[29]</td>
<td>EHH</td>
</tr>
<tr>
<td>KITLG</td>
<td>Non-African</td>
<td>[25, 30-32, 34, 36]</td>
<td>EHH, Tajima’s D, CLR, XP-EHH, iHS, Fst</td>
</tr>
<tr>
<td>LYST</td>
<td>East Asian, European</td>
<td>[30, 34]</td>
<td>Fst, EHH, LSBL, Tajima’s D, lnRH</td>
</tr>
<tr>
<td>MATP</td>
<td>European</td>
<td>[18, 24, 29-31, 34]</td>
<td>Fst, EHH, LSBL, Tajima’s D, lnRH, CLR, LR, FR, XP-EHH</td>
</tr>
<tr>
<td>MITF</td>
<td>All Humans</td>
<td>[25, 29, 30, 34]</td>
<td>EHH, LSBL, Tajima’s D, lnRH, iHS, Tajima’s D</td>
</tr>
<tr>
<td>MYO5A</td>
<td>European</td>
<td>[23, 25]</td>
<td>iHS, EHHS</td>
</tr>
<tr>
<td>OCA2</td>
<td>European, Asian</td>
<td>[18, 25, 29-32, 34, 35]</td>
<td>Fst, EHH, CLR, LSBL, EHHS, iHs, Tajima’s D, FR</td>
</tr>
<tr>
<td>RAB27A</td>
<td>European, East Asian, African</td>
<td>[18, 29, 31]</td>
<td>CLR, EHH, LR, FR</td>
</tr>
<tr>
<td>SILV</td>
<td>East Asian</td>
<td>[31]</td>
<td>CLR</td>
</tr>
<tr>
<td>SLC24A5</td>
<td>European, Middle Eastern, South Asian</td>
<td>[23-25, 29-31, 34, 36]</td>
<td>Fst, EHH, LSBL, Tajima’s D, lnRH, iHS, XP-EHH, CLR, EHHS</td>
</tr>
<tr>
<td>TP53BP1</td>
<td>African</td>
<td>[30]</td>
<td>Fst, EHH</td>
</tr>
<tr>
<td>TYR</td>
<td>East Asian</td>
<td>[29]</td>
<td>EHH</td>
</tr>
<tr>
<td>TYRP1</td>
<td>European</td>
<td>[25, 29, 30, 32, 34, 35]</td>
<td>Fst, EHH, LSBL, EHHS, iHS</td>
</tr>
</tbody>
</table>

All of these studies represent cases where different population samples, different sets of SNPs, and different statistical tests were used in genome-wide screens for selection. That
selection is identified in certain genes in nearly every study (e.g. SLC24A5 and MATP) where other genes are only identified as being under selection in a single study (e.g. EDN3) illustrates the variability that can result due to experimental factors. This is not unique to pigmentation genes as shown by Akey [38] who compiled results from twenty-one whole genome searches for selection. Of the 5,110 genome regions reported as outliers in at least one study, Akey found that only 722 were reported in two studies, 271 in three studies, and 129 in four or more studies. There are several possible reasons why this variability exists including the different time depths explored by the various tests used, the reporting of only the most extreme loci, and low statistical power in addition to the difficulty of accurately calculating power for these types of studies. [38] Because of these inconsistencies, it should not be assumed that results that are not precisely replicated in every study are flawed – though replication certainly bolsters support for selection at the locus.

Among the most striking findings of many of these studies is the magnitude of the selection seen at skin pigmentation genes. The strength of selection on skin pigmentation as populations enter new environments is as strong as or stronger than the selection on genes involved in immunity, reproduction, or diet. Voight et al. [25] found that five pigmentation genes – SLC24A5, OCA2, MYO5A, DTNBPI, and TYRP1 – show a signal of selection in Europeans similar to that seen for the gene LCT which is involved in persistent lactase production allowing adults to digest milk. Myles et al. [39] found that genes for pigmentation were unusually common among selected genes along with genes related to carbohydrate metabolism and skeletal development. Pigmentation genes were more than twice as likely to show evidence of selection than randomly selected genes in both Chinese and European populations in Williamson et al.’s analysis [31]. Pickrell et al. [36] similarly noted that regions of the genome associated with pigmentation are
substantially more likely to have high $F_{st}$ than randomly chosen regions of the genome. Taken together, the large number of pigmentation genes showing strong evidence of selection, including genes such as $EDN3$ which appears to have been selected in common ancestors of all humans [34], suggest a history of selection on pigmentation genes that stretches back to and likely beyond the emergence of the first anatomically modern $Homo sapiens$. Considering the wealth of knowledge about selection at skin pigmentation genes in Old World populations, it is unfortunate to note that no previous researchers have reported on selection at skin pigmentation genes in New World populations as part of a genome-wide investigation. With this background, a research plan was developed to consider from multiple angles the role of selection in shaping skin pigmentation in the Americas.

**Populations and Genotyping Platforms**

Two sets of population samples, genotyped on two different single nucleotide polymorphism (SNP) arrays were used for this study. Incorporating two sets of SNPs and two sets of populations from the same continental regions affords an interesting opportunity to consider the effects of genotyping platform and population choice when trying to recapitulate the evolutionary history of large-scale human migrations. Populations within a single continent – in particular Africa, but also across the whole of the Americas and in Asia – are tremendously variable so one would expect that signatures of selection in the genomes of individuals from distinct, geographically separated populations, such as those included in the Indigenous American samples discussed below, to be markedly different. This could result in the masking of selection events occurring in a single subpopulation which decreases the total amount of information
extracted from the data. However, this masking of unique effects also has the more desirable result of allowing only the strongest and most widespread signatures of selection to show through. Similarly, because the two SNP panels contain relatively few overlapping SNPs, there is a chance that signatures of selection may be detected on one panel but not the other. Considering the broad region of the genome that selection should be altering, it is more likely that when one panel indicates selection, especially in a small number of SNPs, and the other panel does not, there is not strong selection occurring at those loci. So by considering multiple populations and multiple loci, the rate of false-positives should decrease.

Mega Samples

The data set referred to as the “Mega samples” was generated by Abigail Bigham at Pennsylvania State University and Rui Mei at Affymetrix, Inc. (Santa Clara, CA) and are composed of DNA collected from Indigenous American, European American, West African, and East Asian populations. The Indigenous American samples include persons collected from the state of Guerrero, Mexico (Mixtec, Tlapanec, and Nahua speakers, n=14), Maya from the Yucatan Peninsula (n = 25), Quechua from Cerro de Pasco, Peru (n=24), and primarily Aymara with some Quechua individuals from La Paz, Bolivia (n=25). The two South American populations are highland populations which would have been exposed to greater levels of UVR due to the thinner atmosphere at high elevations while the Central American populations are from regions with high levels of UVA similar to those found in West Africa. Although it is impossible to precisely date the amount of time these individuals’ ancestors have inhabited these regions, the inclusion of populations currently residing in high UVR regions in the study affords the best chance for identifying selection favoring darker skin pigmentation. Prior to microarray genotyping, these individuals were genotyped for panels of Ancestry Informative Markers.
(AIMs) to establish that they have extremely low levels of European or other non-Indigenous American ancestry [40]. In addition to the Indigenous American populations, three populations from the HapMap Project were included in the analysis for comparison. These are the Yoruba (Ibadan, Nigeria, n = 60), CEPH Europeans (Utah residents with ancestry from northern and western Europe, n = 60), and East Asians (Han Chinese from Beijing and Japanese from Tokyo, n = 90). These populations are well documented elsewhere [41] but it should be noted that the collection criteria for each of the samples varies. Yoruba donors have four Yoruba grandparents, Han Chinese donors have at least three Han Chinese grandparents, and Japanese donors were simply told the researchers were trying to collect samples from individuals who had ancestors from Japan. The criteria used to collect the CEPH Europeans have not been specified beyond that all the donors are of European ancestry and from Utah. Although the initial sample collections for the Europeans and Yoruba contained parent-offspring trios, only the parents were included in this analysis to avoid confounding due to relatedness. The location of all populations included in the analysis is shown in Figure 2-3 with indications of the amount of ultraviolet A radiation (the primary contributor to folic acid photolysis [42]) reaching the Earth’s surface across all of the world’s land mass.

These samples were genotyped for 906,600 single nucleotide polymorphisms using the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Inc., Santa Clara, CA). The SNPs included on the array are distributed across all of the chromosomes as well as mtDNA with a median autosomal marker spacing of ~1.7kb. The markers included on the Affymetrix array were selected as tagSNPs to capture the variation observed in the HapMap samples and dbSNP (http://www.ncbi.nlm.nih.gov/snp). Because many of these SNPs function to tag the haplotype in which they occur, these tagSNPs represent a greater proportion of the variation in the genome than simply the isolated polymorphism. The Affymetrix array uses a combination of two restriction enzyme-based probes which are discussed in detail on the company’s website and
previously published reviews [43]. Individuals with a genotyping success rate less than 95% were eliminated from analysis.
Human Genome Diversity Project Samples

In addition to the Mega samples, publically available genome-wide SNP data on the HGDP-CEPH samples was analyzed [45]. These individuals were collected by the Human Genome Diversity Project and the Centre d’Étude de Polymorphisme Humain by a number of researchers [46]. Information about the individuals included on the HGDP-CEPH panel is limited to only their sex and population, although previous researchers have also shown that some relative pairs are included in the sample [47]. These individuals were not included in the analysis. Although the geographic distribution is uneven, with enrichment for ethnic groups from China and Pakistan, the sample contains a greater number of populations and more broadly covers continental regions than the Mega sample. The HGDP-CEPH panel consists of 938 individuals in 51 world-wide populations which can be divided into 7 metapopulations following Li et al. [45]: African, Middle Eastern (which includes some northeastern African populations), European, Central-South Asian (predominantly India and Pakistan), East Asian, Oceanian, and American. For consistency with the populations genotyped in the Mega sample, only the 617 individuals from the populations included in the African (n = 121), European (n = 156), East Asian (n = 232), and American (n = 156) groups were used in this analysis. The individual populations that comprise this meta-population and the numbers of individuals typed from each of these populations are listed in Table 2-2. As seen in Figure 2-3, the Indigenous American populations included in the HGDP sample are predominantly localized to the high UVA regions of the Americas. The African populations are similarly clustered in the highest UVA regions of that continent, while the European populations are found in relatively low regions of UVA exposure. The East Asian populations include both high-UVA and low-UVA populations which may make the comparisons with the Indigenous American populations more conservative as the effect of lightening in the high-UVA living East Asian populations may be lessened.
The HGDP samples were genotyped on the Illumina HumanHap650K Beadchips (Illumina, Inc., San Diego, CA) which includes over 650,000 SNPs with a median genomic marker spacing of ~2.0 kb. The Illumina chip uses a series of beads on a microchip array each of which contain a 50bp oligonucleotide to probe a particular SNP marker. [48] The markers were chosen as tagSNPs for the overall variation identified in the HapMap project. The SNPs tag approximately 90% of SNPs in Europeans, 88% of SNPs in East Asians, and 67% of SNPs in Yorubans with minor allele frequencies greater than 5% [45]. Fewer than 200,000 SNPs are found on both the Affymetrix and Illumina arrays. Although this is a relatively small proportion of the total number of SNPs on the arrays, both panels were selected to represent the largest possible amount of overall genome diversity and so we would expect the SNPs to tag most of the same haplotypes within any given gene. It should also be noted that these two platforms are extensions of earlier platforms selected with somewhat different goals in mind. The Affymetrix chip was designed for whole-genome analysis and as such contains SNPs distributed more evenly across the genome. In contrast, the Illumina chip contains more SNPs localized to the genic regions. As a result, the discrepancy between the number of SNPs per gene is less than would be expected considering the total number of SNPs per chip.

Table 2-2. Sample populations and sizes included in meta-populations from HGDP sample.

<table>
<thead>
<tr>
<th>Africans</th>
<th>N = 121</th>
<th>East Asians</th>
<th>N = 232</th>
<th>Europeans</th>
<th>N = 156</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE Bantu</td>
<td>12</td>
<td>Cambodian</td>
<td>11</td>
<td>Adygei</td>
<td>17</td>
</tr>
<tr>
<td>SE Bantu</td>
<td>5</td>
<td>Dai</td>
<td>10</td>
<td>French</td>
<td>29</td>
</tr>
<tr>
<td>SW Bantu</td>
<td>3</td>
<td>Daur</td>
<td>9</td>
<td>Fr Basque</td>
<td>22</td>
</tr>
<tr>
<td>Biaka</td>
<td>32</td>
<td>Han</td>
<td>43</td>
<td>N Italian</td>
<td>13</td>
</tr>
<tr>
<td>Mandenka</td>
<td>24</td>
<td>Hezhen</td>
<td>9</td>
<td>Orcadian</td>
<td>16</td>
</tr>
<tr>
<td>Mbuti</td>
<td>15</td>
<td>Japanese</td>
<td>29</td>
<td>Russian</td>
<td>25</td>
</tr>
<tr>
<td>San</td>
<td>6</td>
<td>Lahu</td>
<td>9</td>
<td>Sardinian</td>
<td>26</td>
</tr>
<tr>
<td>Yoruba</td>
<td>24</td>
<td>Miaoau</td>
<td>10</td>
<td>Tuscan</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mongola</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Naxi</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Americans</td>
<td>N = 108</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colombians</td>
<td>13</td>
<td>Ooroqen</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Karitiana</td>
<td>24</td>
<td>She</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maya</td>
<td>25</td>
<td>Tu</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pima</td>
<td>25</td>
<td>Tujia</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surui</td>
<td>21</td>
<td>Xibo</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yakut</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yizu</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Testing for Selection in Indigenous American Populations

Seventy-six pigmentation candidate genes were investigated for evidence of directional selection unique to Indigenous American populations. These populations were compared with European, West African, and East Asian populations on the basis of three population genetics tests – Locus-Specific Branch Length (LSBL), Reduced Heterozygosity (lnRH), and Tajima’s D – with additional consideration of the size and variation of haplotypes among the populations. The purpose of these tests is to identify regions of the genome which deviate from neutrality in a manner indicative of recent selection in the Indigenous Americans; however, each test measures a somewhat different aspect of the evolutionary process.

The initial list of 76 autosomal pigmentation candidate genes was collected from published sources and the statistical evidence for selection was evaluated for each gene. These genes are ADAM17, ADAMTS20, AP3B1 (ADTB3A), AP3D1, ASIP, ATP7B, ATRN, BCL2, BLOC1S1, BLOC1S3, BNC2, BRCA1, CITED2, CNO, CYP1A2, CYP1B1, CYP2C8, CYP2C9, DAG1, DCT, DTNBP1, ECE1, EDN3, EDNRB, EGFR, ERCC2, FGFR2, GFR3, GNA11, GNAQ, GPNMB, HPS1, HPS3, HPS4, HPS5, HPS6, KIT, KITLG (MGF), KRT1, KRT2A, LMX1A, LYST, MATP (SLC45A2), MC1R, MCOLN3, MGRN1, MITF, MLPH, MREG (DSU), MYO5A, MYO7A, OCA2, OPRM1, PAX3, PLDN, POMC (α-MSH), RAB27A, RAB38, RABGGTA, SFXN1, SILV, SLC24A5, SLC7A11, SNAI2, SNAPAP, SOX10, SOX18, STK11, TBX19 (ACTH), TFAP2A, TYR, TYRP1, VPS13B (COH1, CHS1), VPS33A, WNT1, WNT3A, and ZIC2. In all cases, the statistics were calculated for a region spanning 200,000 bp upstream and downstream of the gene (as defined by the UCSC Genome Browser Build 35 [49]) in an effort to capture near-by regulatory elements. For these analyses, populations were grouped into continental metapopulations as described above.
For each analysis, a broad distribution of values is expected across the genome within each population, representing predominantly those regions of the genome that are evolving neutrally but also those that may be undergoing selection of some kind. Regions of the genome (either genic or regulatory) that are undergoing or have recently undergone selection should be concentrated in the tails of the distributions. The distributions will differ substantially among the tests because the point estimates for particular genes in particular populations are frequently bound to test-specific assumptions and differences in demographic history. When comparing Indigenous American populations to Old World populations, assumptions rooted in similar demographic histories are particularly problematic because of the recent, unique bottleneck or bottlenecks that the Indigenous American populations have experienced. The use of an empirical distribution to rank the values makes the identification of statistical significance relatively model-free because the entire genome would be similarly influenced by founder events, gene flow, or changes in effective population size. In contrast, test values for regions under selection are likely to be distant from the population mean. [50] From an empirical distribution of the values for each SNP or window, those in the top (or bottom, depending on the nature of the test) 5% were considered statistically significant values. It should be noted that when comparing genes of interest, genes that contain more genotyped SNPs will, by chance, contain absolutely more SNPs in the 5% tail, although not proportionally more. In the Affymetrix-genotyped samples there were between 7 and 263 SNPs per gene region with a median number of 50 SNPs. For the smaller Illumina panel, there were between 1 and 204 SNPs per gene region with a median of 30. The number of SNPs in each gene as well as the size and location of each gene can be found in Table 2-3. For either panel, SNPs were eliminated from further analysis if the call rate was less than 95%.

Each of the three tests of selection employed in this study have previously been applied in multiple studies and discussed in detail, but merit review. To identify
directional, as opposed to balancing, selection, these tests locate regions in the genome where there are large genetic distances between populations, low levels of heterozygosity, an excess of rare variants, and regions of extended linkage disequilibrium. Additionally, the tests differ in their ability and means of differentiating between the demographic effects and selection. The different advantages and disadvantages of each test are one important reason for combining them into a single analysis.

**Locus Specific Branch Length**

Directional selection causes an increase in the frequency of a particular allele in a population leading to population differentiation. This can be measured using Locus-Specific Branch Length (LSBL), a test which takes advantage of pairwise measurements of $F_{st}$ from three populations to assign allele frequency changes to a specific population [51]. Developed by Sewall Wright, fixation indices or F-statistics measure the deviation of a population from Hardy-Weinberg equilibrium and can therefore be considered to identify evolution occurring in a population because populations are only in equilibrium when not undergoing evolution. $F_{st}$ [17] describes a special case of deviation due to population subdivisions by estimating the proportion of genetic variation observed in subpopulations relative to the whole. Pairwise $F_{st}$ detects population differentiation by calculating the squared variance ($\sigma$) in allele frequency in a single population divided by the multiplied mean allele frequencies ($p$ and $q$) in the whole population [52].

$$F_{st} = \frac{\sigma_p^2}{\bar{p} \bar{q}}$$
Because large allele frequency differences between two subpopulations requires near fixation of alternate alleles in the populations, \( F_{st} \) is also a measure of reduction in heterozygosity. Thus, \( F_{st} \) can alternately be described by the equation

\[
\frac{H_T - H_S}{H_T}
\]

where \( H_T - H_S \) is the reduction in heterozygosity in the subpopulation compared to the population as a whole. The reliance of \( F_{st} \) on allele frequency carries an implicit assumption of Hardy-Weinberg Equilibrium.

To generate the LSBL values for this analysis, Weir and Cockerham’s unbiased \( F_{st} (\theta) \) [53] was calculated for each pair of populations at each locus. The major advantage of \( \theta \) in contrast with Wright’s \( F_{st} \), is that the former uses the variance to account for the sampling error implicit in the collection of human genetic data (i.e. that it impossible to sample all possible variants in a population). This is accomplished by calculating the proportion of observed variance that is due to between-population variance relative to the total variance as follows:

\[
\hat{\theta} = \frac{a}{a + b + c} = \frac{\text{variance between populations}}{\text{total observed variance}}
\]

where \( a \) is the variance between populations, \( b \) is the variance between individuals within population, and \( c \) is the variance between gametes within individuals. The variables can all be calculated to take sample size directly into account which minimizes the influence of different sample sizes on the analysis.

Large values of \( F_{st} \) indicate a high degree of differentiation between the populations where a value of 1 would be equal to an allele frequency difference between the populations of 1. This means that all of the variance is between the subpopulations. Inversely, small values of \( F_{st} \) indicate that there is little allele frequency differences between the subpopulations with a value of
zero equal to no difference in allele frequency. Small values of $\theta$ are therefore indicate that the majority of the variance is within the subpopulations and the individual. Use of this analysis of variance-based model for $F_{st}$ can result in negative values where the variance is extremely small. Subpopulation differentiation is likely to occur anytime there is substructure causing non-random mating which leads to genetic drift. Therefore, high $F_{st}$ values alone should not be considered sufficient evidence of selection.

Compared to using $F_{st}$ by itself as an indicator of selection, LSBL allows the identification of allele frequency changes unique to the Indigenous American lineage as compared to the East Asian and European populations. While $F_{st}$ will be influenced by fluctuations in allele frequency in any of the populations being compared, LSBL localizes the allele frequency changes to a specific population by triangulation with a third population. Essentially, LSBL creates an individual tree of relatedness for three populations at each locus under investigation. While the pairwise $F_{st}$ values can create a triangle of distance between the three populations, the generated tree is the genetic distance of each population from an inferred common ancestor. The length of each branch on this tree indicates how much that population has changed from the implied ancestor. For each population, branch length can be calculated by summing the two pairwise distances including the population of interest and subtracting the distance between the two reference populations. [51] An illustration and formula for LSBL can be found in Figure 2-4.
Figure 2-4. Diagram and formula for calculating LSBL. LSBL is calculated based on the sum of the genetic distances (D) between the focal population and the other two populations minus the distance between the two non-focal populations divided by half. In this study, LSBL, the value specific to the Indigenous American population is of primary interest and is calculated as LSBL = (d_{IA} + d_{IE} – d_{AE}) / 2.

The most parsimonious explanation for a long branch length in a particular population is that the change has occurred specifically in that lineage. However, this assumes that the three populations diverged simultaneously from the common ancestor. When this is not the case, this attribution of evolution to a particular branch may be inaccurate. For example, a large branch length in Europeans when compared to East Asians and Indigenous Americans could indicate with equal parsimony either that the change occurred in the European population or that it occurred in the common ancestor of East Asians and Indigenous Americans prior to the split between these populations. However, in this analysis, we are interested in identifying allele frequency changes that occurred uniquely in the Indigenous American lineage. A long branch length for the Indigenous American population is a reliable indicator that the evolution occurred in the Indigenous Americans because the alternate explanation – that the same changes
occurred in both Europeans and East Asians after the East Asian and Indigenous American populations split – is substantially less likely. Negative values of LSBL can be achieved when the genetic distance between the two non-focal populations is extremely large due to drift or selection.

Only autosomal SNPs were examined for selection because there are no strong pigmentation genes on either of the sex chromosomes and these require separate analysis from the autosomes because their sex-linked inheritance decreases their effective population size. For each SNP, LSBL\_\text{i} was calculated for the Indigenous American populations with reference to the East Asian and European populations which are their closest relatives from the available samples. LSBL was calculated using a PERL script written by Abigail Bigham [54]. This and all other scripts used in this analysis can be found in Appendix A. SNPs that fell in the top 5% of an empirical distribution of LSBL values for Indigenous Americans were considered to be statistically significant with the large value indicating substantial change in allele frequency in the ancestors of the Indigenous Americans following their split from the other two populations. Stated another way, this indicates substantial population differentiation between the Indigenous Americans and both the East Asians and Europeans. The reason for this one-sided test is that directional selection will reliably increase, never reduce, LSBL in the population undergoing selection. The 5% cutoff value for the Mega sample was 0.2506, while HGDP sample cutoff was lower at 0.2687. The more extreme value for the Mega data is to be expected due to the larger number of SNPs included in the distribution.
**Natural Log Ratio of Heterozygosities (lnRH)**

As discussed, selection will increase the frequency of a particular allele in the population, thereby increasing the expected homozygosity in the population and decreasing the expected heterozygosity or gene diversity. LnRH is a measure of relative gene diversity that identifies regions of the genome where there is reduced heterozygosity compared to a reference population. LnRH was developed [21] based on the earlier lnRV statistic [20] which was a measure of the reduction in variance while lnRH can be computed as follows:

\[
\text{lnRH} = \ln \left[ \frac{E \left( \frac{1}{1-H_{\text{Pop1}}}^2 - 1 \right)}{E \left( \frac{1}{1-H_{\text{Pop2}}}^2 - 1 \right)} \right] = \ln \left( \frac{\text{Gene Diversity in Population 1}}{\text{Gene Diversity in Population 2}} \right)
\]

In this equation, \(H\) is the expected heterozygosity based on the allele frequency in each population. This contributes to the full gene diversity equation for each population. In the absence of selection and drift, the two populations would be expected to maintain the same level of gene diversity found before the populations diverged. Under a neutral model, the expected heterozygosities would be equal in the two populations and lnRH would be 0. Extremely negative values of lnRH across a gene region indicate that there is reduced heterozygosity in Population 1 relative to Population 2. This reduction in heterozygosity may indicate selection in Population 1 due to fixed or nearly fixed allele frequencies.

Reduction in heterozygosity may occur differentially at loci with small minor allele frequencies that become fixed during a bottleneck. To examine the robustness to bottlenecks, Schlötterer and Dieringer ran a series of computer simulations to estimate the effect of a population bottleneck in the focal population while the reference population remained at constant size. LnRH provides a conservative estimate of the degree of reduction in heterozygosity under these parameters. Additionally, even with a small population size and small sample size, the
statistic maintained a normal Gaussian distribution. Although lnRH has been shown to maintain a normal distribution under a broad range of demographic models, all of these models assume independence and the absence of gene flow between the populations under comparison. Because this statistical independence – relatedness or a shared evolutionary history in this context – cannot be true for any two human populations, the expected distribution of the results under a neutral model cannot be known for certain so deviations from the expected distribution are similarly unknowable. This utilization of an empirical distribution of is advantageous because it does not rely on an underlying model of any sort.

LnRH was calculated for each SNP using the Indigenous American/East Asian ratio from a script written by Xianyun Mao and modified by Ellen Quillen to calculate per SNP instead of in windows. Decreased heterozygosity, indicated by low lnRH values, in the Indigenous American populations are suggestive of selection at those loci. Because the fixation of one allele would result in H equaling zero and the statistic being either undefined or zero depending on which population was fixed, allele frequencies of zero were changed to 0.0001 for these calculations. From the distribution of the resultant values, the bottom 5%, representing the most negative values and the most reduced heterozygosity in the Indigenous Americans, was considered significant. The cutoff lnRH value (Indigenous American/East Asian) for the Affy sample was less than -4.2922 and for the HGDP samples less than -5.393.

Tajima’s D

Like lnRH, Tajima’s D can be used to identify deviations from neutrality, but does so by locating regions of the genome with an excess of rare variants. Tajima’s D is calculated based on the ratio of two methods of calculating gene diversity (θ) among a set of sequences – the number of segregating sites (S) and the mean pairwise difference (π) among the sequences [12]. The
number of segregating sites refers to the number of loci that are polymorphic. $\pi$ is calculated by averaging the number of pairwise nucleotide differences between all possible pairs of sequences within the sample. Under neutrality, these two methods of calculating gene diversity should be equal. However, when either directional or balancing selection occurs, the two estimates will differ in predictable ways. The original equation used to calculate Tajima’s D is

$$D = \frac{\pi - S/a_1}{\sqrt{e_1 S + e_2 S(S - 1)}}$$

where $a_1$, $e_1$, and $e_2$, can be derived from the number of sequences in the sample under analysis and incorporate a measure of the standard error of the two estimates of $\theta$. A negative value for $D$ would indicate a larger $S$ which is consistent with both positive selection and population expansion, while a positive $D$ indicates a larger value of $\pi$ which is consistent with both balancing selection and decreasing population size. The difference in these two measures arises because $S$ is more sensitive to the number of low-frequency alleles while $\pi$ is sensitive predominantly to the number of high-frequency alleles. The distribution of frequencies of ancestral and derived alleles is known as the site-frequency spectrum. After the diversity in a region is eliminated by directional selection, new neutral mutations will arise and generally be eliminated by drift or remain at low frequency in the population. Because so few individuals in the sample will harbor these rare variants, they increase the total number of sites at which there are polymorphisms (inflating the frequency of $S$) to a greater degree than they increase the number of pairwise differences ($\pi$) which will result in a negative value of Tajima’s D. While initially developed for the analysis of DNA sequence data, Tajima’s D has been extended to analyze dense SNP data. The predominant concern with the use of dense SNP data is that the inclusion of only relatively common polymorphic sites will bias the values of Tajima’s D upwards. Since this makes the test more conservative for directional selection, this is not a major concern.
One limitation regarding Tajima’s D is that the ability to detect selection based on an excess of rare variants relies on the assumption that a new allele will rise to fixation quickly thereby eliminating variation in a broad surrounding region. If it is instead the case that the environment changes and an existing allele becomes favored, the ability of Tajima’s D to identify selection is diminished [56]. Due to recombination, an existing variant will likely be found in multiple haplotypic backgrounds so that when selection occurs, the sequence surrounding the gene will remain polymorphic even when the adaptive allele has become fixed. If the adaptive allele was slightly advantageous or neutral in the previous environment, then it could have been maintained at a high enough frequency to diminish the breadth of the selective sweep that must occur prior to the rebound in variation detected by Tajima’s D. If the alleles existed in the ancestral population but were maintained a low frequency due to negative selection or drift, then the probability of detecting the selection event is increased.

Because the frequency of segregating sites can only be calculated for a series of SNPs, Tajima’s D is calculated for sliding windows 10,000 base pairs in length at 5,000 base pair intervals. As a result of the uneven distribution of SNPs across the genome, there are different numbers of SNPs in each window. Windows with fewer than 5 successfully genotyped SNPs were not analyzed for Tajima’s D.

One concern with examining Tajima’s D in a population like Indigenous Americans is that the bottleneck is usually followed by a rapid population expansion during which time a large number of rare alleles are generated in the population (commonly referred to a star-like phylogeny). [57] To address this issue, Tajima’s D was normalized by dividing the D calculated for each window by the genome-wide average D for each population. This reduces the total number of genes with extreme (greater than two standard deviations from the mean) values in Indigenous Americans with the assumption that many of these extreme values are based on demographic effects alone. This normalization is an important first step before comparing the D
values among populations because of the different demographic histories of the populations. Normalized Tajima’s D significance cutoff value were less than -1.9004 for the Affy samples and -1.8923 for the HGDP samples. Because the focus of this research is directional selection, only the negative tail of Tajima’s D was considered but the positive tail could also be considered for investigations of balancing selection.

To better differentiate between of changes that are unique to the Indigenous American populations instead of carried over effects of selection in the common Central Asian ancestral population, Tajima’s D Difference was calculated, following Bigham et al. [54]. Tajima’s D Difference is calculated for each window with normalization as

$$Tajima's \text{ D Difference} = \frac{(D_I - D_A) - \mu(D_I - D_A)}{SD (D_I - D_A)}$$

where $D_I$ and $D_A$ are the Tajima’s D values for the Indigenous American and East Asian populations, $\mu$ is the mean D value across all windows, and SD is the standard deviation of the difference. This allows for the identification Where there is a negative Tajima’s D value for the Indigenous Americans but a positive (or weakly negative) D for the East Asians, the Tajima’s D Difference will remain negative indicating that there are different selective pressures at work in the two populations. In contrast, where the same selection has been occurring in both the Indigenous American and East Asian populations, the Tajima’s D Difference value will become more positive, eliminating that gene region from the negative significance tail. For Tajima’s D Difference, the lower 5% tail cutoff was -1.6695 in the Mega sample and -3.2550 in the HGDP samples. These more extreme values for the D Difference are indicative of a broader distribution of values, pushed lower by cases were there were negative D values in the Indigenous Americans and positive D values in the East Asians. All Tajima’s D statistics were calculated from a PERL script written by Xianyun Mao and modified by Abigail Bigham.
Assessment of Haplotype Structure

While not a statistical test, a final means of investigating the data is to consider the haplotypes block structure in the Indigenous Americans to identify regions of extended linkage disequilibrium (LD) in the genome. That regions of LD can be generated by selection has been previously discussed. Mapping the global variation of these LD blocks was one of the primary goals at the inception of the International HapMap project, in part to look for regions of the genome that may have undergone selection [58]. To visualize the haplotypes blocks, the SNP data was first phased to create haplotypes using fastPHASE [59]. This program uses a hidden Markov model to impute missing genotypes and estimate haplotypes structure along the genome. This haplotypes were then analyzed using Haploview [60] which can calculate a variety of statistics to quantify linkage disequilibrium. Hedrick’s multiallelic $D'$ was the measure used in this study [61]. This is among the most commonly used measures of LD and is expected to show the largest values at the core of the haplotypes which should be located near the focus of selection and decrease with increasing distance from that locus and depending on local recombination rates [62]. Haploview plots the reach of LD based on the $D'$ values, providing a visual representation of the extent and pattern of the LD blocks in different populations. As large-scale data sets became available, LD maps were generated for many populations. Among the most surprising findings was the large number of haplotypes that were shared among populations, particularly non-African populations. [63] This sharing supports the idea that genomic regions showing radically dissimilar, extended haplotype structures may be indicative of selection. Divergent patterns between Indigenous Americans and East Asians could be a signature of selection, but it essential to consider that genetic bottlenecks may also increase both the length and frequency of particular haplotypes due to an overall reduction in genetic variation. [64] Because the
Indigenous American populations have been subject to recent and severe bottlenecks, LD blocks may be extended due to demographic effects.

**Necessity of Incorporating Multiple Tests**

The inclusion of multiple tests of selection in these analyses has the two-fold benefit of decreasing both the overall false-positive and false-negative rates. Because the tests are all measures of selection, evidence of selection in only one test indicates a strong likelihood that the single test with a significant result may simply be due to the arbitrary 5% cut off and not a true signal that the gene has undergone recent directional selection. However, the likelihood that the same neutral region of the genome would by chance show statistically significant results on multiple tests – and particularly across both samples – is small.

Additionally, incorporating multiple tests of selection diminishes the likelihood that demographic histories – most saliently, the recent bottleneck as Indigenous American populations migrated into the New World – will create false positives because the tests are influenced differently by demography. Tajima’s D in particular can be sensitive to recent bottlenecks which may increase the overall number of extreme values. However, the normalization and comparison of two populations should account for this. LnRH has performed well under a broad range of demographic scenarios including bottlenecks but the precise suitability of this test to the still incompletely resolved demographic history of the Americas is not well documented [21]. LSBL, because it is based on $F_{st}$, is extremely sensitive to demographic parameters because it responds to any cause of population differentiation including drift. The assumption is that selection will cause some proportion of the change in allele frequencies among the populations analyzed and that those SNPs with the highest $F_{st}$ values are those most likely to have been subject to selection.
However, there is no other way to differentiate between large values caused by demography or selection.

Finally, it is important to consider the limitations of each test for assessing different selection regimes and how the tests compensate for the limitations of others. For example, Tajima’s D measures the increase in rare alleles following a selective sweep. In cases where the selective pressure is ongoing or too old, Tajima’s D will be inefficient. Similarly, if the selected mutation exists in several haplotypes (likely because the polymorphism pre-exists the onset of selection), Tajima’s D values will be less extreme [65]. Incomplete sweeps are also a challenge to detect using Tajima’s D because of the large amount of variation that may still remain in the region following a partial sweep.

Tests such as lnRH which rely on the identification of regions of reduced heterozygosity must be interpreted in light of the large size of genomic regions which may show reduction in heterozygosity and the long period of time over which this reduction may persist. A broad region of reduced heterozygosity makes localization of the gene under selection difficult because there may be many genes included in the more than 600 kb region of reduced heterozygosity that may result from a selective advantage of as little as 1% [66]. This sensitivity can compensate for some of the problems with Tajima’s D in detecting incomplete sweeps. It cannot be assumed that the selected gene must be in the center of the selected region because recombination rates vary across the genome and recombination hotspots may erase evidence of prior selection. The combination of these tests is useful in narrowing down the region of the genome that has undergone selection. In particular, LSBL generally indicates a narrower genomic region than the other tests [67]. By including multiple tests, the regions where multiple tests show evidence for selection are more likely to be the portions of the gene or regulatory region that have undergone selection. The duration of the signature of selection is an additional concern because evidence detected in the Indigenous American populations may be a remnant of selection in ancestral populations. By
comparing the ratio of heterozygosity in the Indigenous Americans and East Asians, this concern is reduced.

LSBL will be better able to identify selection that has occurred on pre-existing variation than Tajima’s D or lnRH because it considers the allele frequency differences between the two populations at a single locus rather than across the region. Where selection acts to favor a polymorphism that already exists at low frequencies in multiple backgrounds, there will be an increased allele frequency difference, and therefore an increased LSBL value, but the variation in the surrounding region will not be substantially reduced and so will be invisible to lnRH and Tajima’s D. This is because lnRH fails if the locus is near fixation prior to selection.

Power to Detect Selection

No selection-based studies have reported the statistical power of their tests because calculating power and the expected false positive rates require accurate estimation of a number of parameters that are impossible to know a priori if at all. These include the proportion of the genome that is undergoing and has undergone selection, the strength of selection, the degree of polymorphism at the locus prior to the onset of selection, and whether the favored allele is dominant, recessive or additive. In particular, the false-positive rate can be substantially inflated when only a small percentage of genes are under selection or if selection is generally weak. [68]

When researchers validate new tests, most report on the ability of the tests to identify genes that are previously known to have undergone selection and/or the projected ability of their method to detect selection in a variety of simulated data sets. This is the case with all of the papers that first report on the tests of selection used in this research. None of these offer a means of estimating statistical power. Although a numerical power calculation cannot be done, it is expected that the largest possible sample size will yield the best results. When employing an
empirical distribution and assuming that the most strongly selected regions will fall in the outlying region, then tests will have more power when the sample variance of the statistic is smaller. [21]

To test the efficacy of six empirically-based tests of selection including Tajima’s D, Teshima et al. [69] generated an artificial data set based on a simplified version of a domesticated crop (maize) and human demographic history with directional selection of varying strengths, with randomly chosen levels of dominance, and the ability for selection to act either on a novel or standing genetic variation. By definition, precise calculations of power and false-discovery rates cannot be done for an empirically based test because it is model-free.

Using the maize model, Teshima et al. found that new mutations that are rapidly swept to fixation are frequent in the tails of empirical distributions of multiple tests of selection. As can be seen in Figure 2-5, the false discovery and false negative rates are extremely small for new mutations where the overall frequency of selection is high genome-wide. When the frequency ($f$) of an allele is 0.05 prior to the onset of selection, both the false discovery and, to a greater extent, the false negative rate increase markedly. At a 5% significance cutoff where the proportion of loci under selection is low genome-wide, the false discovery and false negative rates can reach up to 86% and 94% for Tajima’s D. This indicates that it is nearly impossible to differentiate between neutrally evolving alleles and alleles that have been selected from standing variation.
Figure 2-5. Estimated False Discovery or False Negative Rates using Empirical Models. The statistical cutoff used to measure significance (x axis), proportion of loci undergoing selection (y axis), and whether the adaptive alleles is a new mutation or part of standing variation will all influence the false discovery and false negative rates. Differentiating between neutrally evolving and selected alleles is substantially easier when the favored allele is newly evolved and cutoff is approximately equal to the proportion of selected loci in the genome. Unfortunately, there is no a priori way to determine the proportion of loci undergoing selection or whether that selection is on new or existing polymorphisms. Figure from [69].

This would seem to deal a heavy blow against the use of empirical distributions to detect selection on standing variation. However, the simulation of human data revealed an interesting component to this puzzle. When selection on standing variation follows a bottleneck – as would be expected to happen when the earliest Americans expanded into equatorial regions of the New World, the ability to detect this selection increases. This is likely due to the overall reduction in variation genome-wide which will cause the variability of the genetic background of the pre-existing allele to be similar to what would be expected for a newly emerging allele. When the selection is modeled as being simultaneous to the bottleneck, the existing allele frequency once again influences the power of the tests.
The inadequate ability to estimate power for these analyses illustrates the necessity of further investigation of genes showing evidence of selection to establish or disprove links between the genes and phenotypic variation. The results discussed below are important in that they narrow down an intractably large number of potentially functional pigmentation genes into a more manageable number of selection-nominated candidate genes. False-negatives and false-positives are both major issues when dealing with an empirical distribution so considering overall patterns rather than classifying genes as significant or non-significant is a major topic covered in the following section.

**Results of Tests for Selection**

Fifty-five of the 76 candidate genes tested have top 5\% significance values for at least one of these tests in one SNP or window within the gene region but due to the large number of tests being performed, many of these results are likely to be anomalous. From this set of genes, only those genes are considered that show evidence in multiple windows/SNPs and for multiple tests. The distribution of significant values for the three tests of selection across all of the candidate genes can be seen in Table 2-3. To establish what distribution of results would be expected for genes that have undergone selection, the phenomenon of genetic hitchhiking must be considered. Because recombination is a relatively slow process in the human genome, linkage disequilibrium will occur among genes near one another on a chromosome. This phenomenon, where genes are inherited together as a unit, leads to a broad region of the genome showing evidence of selection with the strongest evidence focused on the functional region and decreasing evidence with increasing distance from the locus under selection. [70] The magnitude of this effect is heavily dependent on local recombination rates because recombination hotspots can mask the effect of hitchhiking by frequently introducing new variation. [37] A pattern of broad
regions of selection with a peak near the gene is what is considered strong evidence when considering the results of the statistical analyses.

Therefore, looking at the overall pattern of significant results within genes and across tests generates a more accurate picture of the effect of selection on the gene. For example, a broad region of positive LSBL values that do not reach significance may be better evidence for selection than a single, highly significant value. Likewise, similar patterns of statistical values across the different tests may be a better indicator of selection (and accurate results) than a few significant values on only a single test. The implication of selection in this way is somewhat more arbitrary than simply identifying those SNPs with the most extreme values but hews more closely to theoretical expectations of the effects of selection on the genome. A smaller, broader effect is more likely to represent a selection event than a narrow, very extreme one which could be a chance outlier, a genotyping error, or some other anomaly. The reasons for selecting particular pigmentation genes for further study are covered for each gene in the next section.
Table 2-3. Percentages of SNPs/windows showing statistically significant results for each test in each population. This table details the numerical result of each test averaged across each gene region as well as the percentages (in parenthesis) of SNPs (or windows in the case of Tajima’s D) in each gene region found in the 5% significance tail. Due to variations in the density of SNPs between the two genotyping platforms and among the genes, gene size is not a perfect proxy of number of SNPs so the number of SNPs present in each gene region on each platform is indicated. Tajima’s D as indicated here is the normalized Tajima’s D difference as discussed in the text.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>Location (kb)</th>
<th>SNPs</th>
<th>LSL</th>
<th>lnR⁹</th>
<th>Tajima’s D</th>
<th>SNPs</th>
<th>LSL</th>
<th>lnR⁹</th>
<th>Tajima’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECE1</td>
<td>1</td>
<td>21419-21479</td>
<td>45</td>
<td>0.0132 (2.2%)</td>
<td>0.7823 (4.4%)</td>
<td>-0.2712 (0.0%)</td>
<td>49</td>
<td>-0.0063 (0.0%)</td>
<td>-0.1223 (0.0%)</td>
<td>-1.0938 (0.0%)</td>
</tr>
<tr>
<td>MCOLN3</td>
<td>1</td>
<td>85263-85287</td>
<td>42</td>
<td>0.0332 (0.0%)</td>
<td>-0.3105 (2.4%)</td>
<td>-0.2874 (0.0%)</td>
<td>27</td>
<td>0.0262 (0.0%)</td>
<td>0.0987 (0.0%)</td>
<td>-1.7585 (0.0%)</td>
</tr>
<tr>
<td>SNAPAP</td>
<td>1</td>
<td>151898-151901</td>
<td>15</td>
<td>0.0686 (0.0%)</td>
<td>0.1263 (0.0%)</td>
<td>-1.2150 (0.0%)</td>
<td>11</td>
<td>0.1418 (9.1%)</td>
<td>-1.3125 (0.0%)</td>
<td>-1.8578 (0.0%)</td>
</tr>
<tr>
<td>LMX1A</td>
<td>1</td>
<td>163438-163592</td>
<td>168</td>
<td>0.0140 (0.0%)</td>
<td>0.7166 (3.6%)</td>
<td>0.0187 (0.0%)</td>
<td>107</td>
<td>0.0361 (0.9%)</td>
<td>-0.4366 (5.6%)</td>
<td>-1.1958 (0.0%)</td>
</tr>
<tr>
<td>TBX19</td>
<td>1</td>
<td>166533-166550</td>
<td>50</td>
<td>0.0274 (2.0%)</td>
<td>-0.2297 (2.0%)</td>
<td>-0.1779 (0.0%)</td>
<td>29</td>
<td>0.0334 (0.0%)</td>
<td>-0.6468 (3.4%)</td>
<td>-1.0810 (0.0%)</td>
</tr>
<tr>
<td>WNT3A</td>
<td>1</td>
<td>226261-226316</td>
<td>19</td>
<td>0.0237 (0.0%)</td>
<td>-0.3057 (5.3%)</td>
<td>0.1770 (0.0%)</td>
<td>15</td>
<td>0.0017 (0.0%)</td>
<td>-0.0285 (0.0%)</td>
<td>-1.6526 (0.0%)</td>
</tr>
<tr>
<td>LYST</td>
<td>1</td>
<td>234059-234097</td>
<td>31</td>
<td>-0.0013 (0.0%)</td>
<td>1.8023 (0.0%)</td>
<td>0.5717 (0.0%)</td>
<td>13</td>
<td>-0.0246 (0.0%)</td>
<td>1.5117 (0.0%)</td>
<td>-1.7030 (0.0%)</td>
</tr>
<tr>
<td>ADAM17</td>
<td>2</td>
<td>9585-9613</td>
<td>34</td>
<td>-0.0015 (0.0%)</td>
<td>0.0470 (8.8%)</td>
<td>-0.1858 (0.0%)</td>
<td>13</td>
<td>0.0542 (0.0%)</td>
<td>-2.2815 (15.4%)</td>
<td>-2.3519 (20.0%)</td>
</tr>
<tr>
<td>POMC</td>
<td>2</td>
<td>25237-25245</td>
<td>26</td>
<td>0.1667 (30.8%)</td>
<td>-0.6437 (3.8%)</td>
<td>-0.1890 (0.0%)</td>
<td>16</td>
<td>0.2140 (43.8%)</td>
<td>-0.9387 (6.3%)</td>
<td>-0.7978 (0.0%)</td>
</tr>
<tr>
<td>CYP1B1</td>
<td>2</td>
<td>38148-38157</td>
<td>69</td>
<td>-0.0123 (0.0%)</td>
<td>0.4181 (7.2%)</td>
<td>1.2694 (0.0%)</td>
<td>41</td>
<td>0.0201 (0.0%)</td>
<td>-0.8517 (9.8%)</td>
<td>-1.0013 (0.0%)</td>
</tr>
<tr>
<td>MREG</td>
<td>2</td>
<td>216516-216587</td>
<td>74</td>
<td>0.0238 (0.0%)</td>
<td>-0.1018 (0.0%)</td>
<td>-0.7427 (0.0%)</td>
<td>66</td>
<td>0.0559 (9.1%)</td>
<td>-0.6664 (0.0%)</td>
<td>-1.3078 (0.0%)</td>
</tr>
<tr>
<td>PAX3</td>
<td>2</td>
<td>222773-222872</td>
<td>54</td>
<td>0.0631 (3.7%)</td>
<td>0.3592 (3.7%)</td>
<td>-1.8247 (75.0%)</td>
<td>56</td>
<td>0.0720 (8.9%)</td>
<td>-1.4112 (10.7%)</td>
<td>-2.1964 (25.0%)</td>
</tr>
<tr>
<td>MLPH</td>
<td>2</td>
<td>238060-238129</td>
<td>78</td>
<td>0.0278 (0.0%)</td>
<td>-0.4414 (7.7%)</td>
<td>0.4170 (0.0%)</td>
<td>66</td>
<td>0.0338 (0.0%)</td>
<td>-0.8978 (3.0%)</td>
<td>-2.5279 (0.0%)</td>
</tr>
<tr>
<td>DAI1</td>
<td>3</td>
<td>49483-49548</td>
<td>11</td>
<td>0.0514 (0.0%)</td>
<td>-2.2786 (18.2%)</td>
<td>-0.1073 (0.0%)</td>
<td>8</td>
<td>0.0870 (0.0%)</td>
<td>-1.8518 (0.0%)</td>
<td>-1.4679 (0.0%)</td>
</tr>
<tr>
<td>MTF</td>
<td>3</td>
<td>70068-71000</td>
<td>24</td>
<td>0.0524 (0.0%)</td>
<td>-1.1752 (12.5%)</td>
<td>-0.4015 (0.0%)</td>
<td>8</td>
<td>0.0218 (0.0%)</td>
<td>-0.8699 (12.5%)</td>
<td>-0.9167 (0.0%)</td>
</tr>
<tr>
<td>HPS3</td>
<td>3</td>
<td>150330-150374</td>
<td>38</td>
<td>0.0391 (0.0%)</td>
<td>0.7700 (2.6%)</td>
<td>-0.7122 (0.0%)</td>
<td>32</td>
<td>0.0217 (0.0%)</td>
<td>0.2552 (3.1%)</td>
<td>-2.3372 (16.7%)</td>
</tr>
<tr>
<td>CNO</td>
<td>4</td>
<td>6769-6770</td>
<td>31</td>
<td>0.0379 (0.0%)</td>
<td>0.1270 (0.0%)</td>
<td>0.3803 (0.0%)</td>
<td>40</td>
<td>0.0592 (5.0%)</td>
<td>-0.6430 (5.0%)</td>
<td>-1.0108 (0.0%)</td>
</tr>
<tr>
<td>KIT</td>
<td>4</td>
<td>55219-55302</td>
<td>77</td>
<td>0.0478 (0.0%)</td>
<td>-0.8750 (9.1%)</td>
<td>-0.4803 (14.3%)</td>
<td>62</td>
<td>0.0477 (1.6%)</td>
<td>-0.8032 (6.5%)</td>
<td>-2.3558 (14.3%)</td>
</tr>
<tr>
<td>SLC7A11</td>
<td>4</td>
<td>139305-139383</td>
<td>42</td>
<td>0.0074 (0.0%)</td>
<td>0.6325 (2.4%)</td>
<td>0.1159 (0.0%)</td>
<td>27</td>
<td>0.0075 (0.0%)</td>
<td>-0.1406 (3.7%)</td>
<td>-1.8316 (0.0%)</td>
</tr>
<tr>
<td>MATP</td>
<td>5</td>
<td>33980-34021</td>
<td>70</td>
<td>0.0568 (0.0%)</td>
<td>-0.6140 (0.0%)</td>
<td>-0.3984 (0.0%)</td>
<td>55</td>
<td>0.0756 (5.5%)</td>
<td>-0.4753 (0.0%)</td>
<td>-0.3733 (0.0%)</td>
</tr>
<tr>
<td>Gene</td>
<td>Value 1</td>
<td>Value 2</td>
<td>Value 3</td>
<td>Value 4</td>
<td>Value 5</td>
<td>Value 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP3B1</td>
<td>0.0989</td>
<td>0.3156</td>
<td>0.6980</td>
<td>0.1065</td>
<td>0.2752</td>
<td>-2.5898</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFXN1</td>
<td>0.0106</td>
<td>-0.2560</td>
<td>-0.0955</td>
<td>0.0151</td>
<td>-0.1157</td>
<td>-2.2880</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFFAP2</td>
<td>0.1475</td>
<td>-0.6366</td>
<td>0.2148</td>
<td>0.1029</td>
<td>-0.6578</td>
<td>-1.6778</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTNBP1</td>
<td>0.0293</td>
<td>0.9859</td>
<td>0.2940</td>
<td>0.0220</td>
<td>0.1527</td>
<td>-1.0928</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CITED2</td>
<td>0.0060</td>
<td>1.0006</td>
<td>0.1101</td>
<td>0.0161</td>
<td>-0.1276</td>
<td>-1.2231</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPRM1</td>
<td>0.1381</td>
<td>0.0492</td>
<td>-0.5017</td>
<td>0.1594</td>
<td>-0.1880</td>
<td>-2.4862</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPNMB</td>
<td>0.0649</td>
<td>0.2429</td>
<td>0.0445</td>
<td>0.0646</td>
<td>-0.1746</td>
<td>-1.1762</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGFR</td>
<td>0.0204</td>
<td>1.1850</td>
<td>0.4899</td>
<td>0.0346</td>
<td>0.2144</td>
<td>-2.8200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAI2</td>
<td>0.0424</td>
<td>1.2351</td>
<td>0.6832</td>
<td>0.0929</td>
<td>-0.6610</td>
<td>-0.4250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPS13B</td>
<td>0.1331</td>
<td>-0.3213</td>
<td>-0.0565</td>
<td>0.0737</td>
<td>-0.3309</td>
<td>-0.4895</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TYRP1</td>
<td>-0.0189</td>
<td>1.4593</td>
<td>1.1912</td>
<td>-0.0004</td>
<td>-0.1562</td>
<td>-1.3865</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNC2</td>
<td>0.0398</td>
<td>0.3813</td>
<td>0.2495</td>
<td>0.0473</td>
<td>0.0980</td>
<td>-1.6749</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNAQ</td>
<td>0.0282</td>
<td>0.1045</td>
<td>-0.4698</td>
<td>0.0428</td>
<td>-0.6354</td>
<td>-1.7942</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2C9</td>
<td>0.0882</td>
<td>-1.7212</td>
<td>-1.9591</td>
<td>0.1095</td>
<td>-2.3534</td>
<td>-1.4921</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2C8</td>
<td>0.1379</td>
<td>-0.5929</td>
<td>-0.6308</td>
<td>0.0973</td>
<td>-1.4972</td>
<td>-1.4409</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPS1</td>
<td>0.0507</td>
<td>-0.2491</td>
<td>-0.3954</td>
<td>0.0675</td>
<td>-0.8556</td>
<td>-2.1609</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPS6</td>
<td>0.0417</td>
<td>2.3772</td>
<td>-1.1488</td>
<td>1</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGFR2</td>
<td>0.0603</td>
<td>0.2905</td>
<td>0.1365</td>
<td>0.0962</td>
<td>-0.4081</td>
<td>-1.4988</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPS5</td>
<td>-0.0130</td>
<td>0.7323</td>
<td>0.2374</td>
<td>0.0017</td>
<td>-0.2408</td>
<td>-1.4580</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYO7A</td>
<td>0.0498</td>
<td>-0.4208</td>
<td>-1.2650</td>
<td>0.0477</td>
<td>-0.3156</td>
<td>-2.5200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAB38</td>
<td>0.0154</td>
<td>-0.2111</td>
<td>0.1939</td>
<td>0.0037</td>
<td>-0.5552</td>
<td>-1.8682</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TYR</td>
<td>0.0095</td>
<td>0.2816</td>
<td>-0.1688</td>
<td>0.0127</td>
<td>-0.4542</td>
<td>-1.8912</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADAMTS20</td>
<td>0.0045</td>
<td>-0.1652</td>
<td>0.8019</td>
<td>0.0222</td>
<td>-0.9852</td>
<td>-1.5838</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WNT1</td>
<td>0.0584</td>
<td>-0.3555</td>
<td>0.3082</td>
<td>0.0472</td>
<td>-0.8749</td>
<td>-2.3540</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRT2A</td>
<td>0.0841</td>
<td>0.2079</td>
<td>0.7159</td>
<td>0.0572</td>
<td>-0.8699</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRT1</td>
<td>0.0690</td>
<td>0.6177</td>
<td>0.9163</td>
<td>0.1130</td>
<td>-0.5497</td>
<td>-2.0036</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLOC1S1</td>
<td>0.0243</td>
<td>0.3364</td>
<td>1.2652</td>
<td>0.0804</td>
<td>-0.5078</td>
<td>-1.6243</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SILV</td>
<td>0.0634</td>
<td>3.0086</td>
<td>-2.6860</td>
<td>0.0655</td>
<td>1.2742</td>
<td>-1.4607</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Value</td>
<td>Fold Change</td>
<td>p-Value</td>
<td>q-Value</td>
<td>Adj Fold Change</td>
<td>p-Value</td>
<td>q-Value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-----------</td>
<td>-------------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------------</td>
<td>-----------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KITLG</td>
<td>12</td>
<td>87411-87498</td>
<td>0.0606 (0.0%)</td>
<td>-3.0323 (27.9%)</td>
<td>-2.4460 (100.0%)</td>
<td>0.0375 (0.0%)</td>
<td>-2.3531 (21.7%)</td>
<td>-0.6550 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPS33A</td>
<td>12</td>
<td>121258-121317</td>
<td>0.1023 (0.0%)</td>
<td>0.0361 (4.5%)</td>
<td>0.5762 (0.0%)</td>
<td>0.1449 (15.8%)</td>
<td>-0.2129 (0.0%)</td>
<td>-1.3105 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP7B</td>
<td>13</td>
<td>51405-51484</td>
<td>0.0412 (0.0%)</td>
<td>-0.1498 (2.2%)</td>
<td>-0.2766 (0.0%)</td>
<td>0.0355 (0.0%)</td>
<td>0.0465 (0.0%)</td>
<td>-0.4594 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDN3</td>
<td>13</td>
<td>77368-77448</td>
<td>0.0032 (0.0%)</td>
<td>-0.4428 (10.2%)</td>
<td>0.3247 (0.0%)</td>
<td>0.0193 (0.0%)</td>
<td>-1.4185 (0.0%)</td>
<td>-1.3671 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCT</td>
<td>13</td>
<td>93890-93930</td>
<td>-0.0331 (0.0%)</td>
<td>0.7993 (0.0%)</td>
<td>0.2209 (0.0%)</td>
<td>0.0373 (0.0%)</td>
<td>-1.2340 (12.5%)</td>
<td>-1.2660 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZIC2</td>
<td>13</td>
<td>99432-99437</td>
<td>0.0235 (0.0%)</td>
<td>0.7522 (4.5%)</td>
<td>-0.5839 (0.0%)</td>
<td>0.0392 (0.0%)</td>
<td>-0.4890 (4.2%)</td>
<td>-1.6221 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RABGTA</td>
<td>14</td>
<td>23805-23811</td>
<td>0.0764 (8.3%)</td>
<td>0.3975 (0.0%)</td>
<td>-0.5464 (0.0%)</td>
<td>0.0475 (0.0%)</td>
<td>-0.2481 (5.6%)</td>
<td>-1.2051 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OCA2</td>
<td>15</td>
<td>25674-26018</td>
<td>-0.0009 (0.0%)</td>
<td>1.2730 (4.0%)</td>
<td>0.8699 (0.0%)</td>
<td>0.0097 (0.8%)</td>
<td>0.1531 (1.6%)</td>
<td>-0.5106 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLDN</td>
<td>15</td>
<td>43667-43689</td>
<td>0.2184 (37.0%)</td>
<td>-0.9071 (3.7%)</td>
<td>-1.3353 (0.0%)</td>
<td>0.2042 (39.1%)</td>
<td>-1.7813 (8.7%)</td>
<td>-2.4484 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLC2A5</td>
<td>15</td>
<td>46200-46222</td>
<td>0.0003 (0.0%)</td>
<td>0.1458 (4.8%)</td>
<td>-0.7142 (20.0%)</td>
<td>0.0757 (33.3%)</td>
<td>0.8660 (0.0%)</td>
<td>-1.1239 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYO5A</td>
<td>15</td>
<td>50387-50609</td>
<td>0.0171 (0.0%)</td>
<td>-0.4558 (5.2%)</td>
<td>-0.8009 (30.8%)</td>
<td>0.0063 (0.0%)</td>
<td>-1.2172 (13.3%)</td>
<td>-1.5372 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAB27A</td>
<td>15</td>
<td>53283-53369</td>
<td>0.0810 (9.8%)</td>
<td>0.1236 (2.0%)</td>
<td>0.2648 (0.0%)</td>
<td>0.0854 (2.7%)</td>
<td>-2.0188 (24.3%)</td>
<td>-1.5704 (14.3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP1A2</td>
<td>15</td>
<td>72828-72836</td>
<td>0.0837 (0.0%)</td>
<td>0.2921 (0.0%)</td>
<td>-0.3816 (0.0%)</td>
<td>0.1202 (11.8%)</td>
<td>-0.7463 (5.9%)</td>
<td>-2.5927 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGRN1</td>
<td>16</td>
<td>4615-4681</td>
<td>0.0305 (0.0%)</td>
<td>0.0724 (11.1%)</td>
<td>0.6877 (0.0%)</td>
<td>0.0154 (0.0%)</td>
<td>0.1433 (0.0%)</td>
<td>-0.8769 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCTR</td>
<td>16</td>
<td>88513-88515</td>
<td>0.0666 (0.0%)</td>
<td>-2.9614 (7.7%)</td>
<td>-0.3430 (25.0%)</td>
<td>0.0383 (0.0%)</td>
<td>-1.3844 (8.3%)</td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA1</td>
<td>17</td>
<td>38450-38531</td>
<td>0.0011 (0.0%)</td>
<td>0.6631 (0.0%)</td>
<td>0.3585 (0.0%)</td>
<td>0.0352 (0.0%)</td>
<td>-0.5828 (10.0%)</td>
<td>-1.3800 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCL2</td>
<td>18</td>
<td>58942-59138</td>
<td>0.0014 (0.0%)</td>
<td>0.4409 (2.7%)</td>
<td>0.3938 (0.0%)</td>
<td>0.0456 (1.9%)</td>
<td>-0.6376 (2.8%)</td>
<td>-1.4670 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STK11</td>
<td>19</td>
<td>1157-1179</td>
<td>-0.0196 (0.0%)</td>
<td>1.3707 (0.0%)</td>
<td>0.9744 (0.0%)</td>
<td>0.0396 (0.0%)</td>
<td>-0.4432 (0.0%)</td>
<td>-0.7743 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP3D1</td>
<td>19</td>
<td>2052-2103</td>
<td>0.0301 (0.0%)</td>
<td>-0.2329 (0.0%)</td>
<td>-0.4504 (0.0%)</td>
<td>0.0902 (6.7%)</td>
<td>-1.1414 (10.0%)</td>
<td>-2.5360 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNA11</td>
<td>19</td>
<td>3045-3072</td>
<td>0.0414 (0.0%)</td>
<td>-0.2442 (0.0%)</td>
<td>0.1720 (0.0%)</td>
<td>0.1171 (13.3%)</td>
<td>-1.1287 (3.3%)</td>
<td>-1.4872 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLOC1S3</td>
<td>19</td>
<td>50374-50375</td>
<td>0.0101 (0.0%)</td>
<td>0.6407 (0.0%)</td>
<td>-0.2821 (25.0%)</td>
<td>0.0403 (0.0%)</td>
<td>-0.7941 (0.0%)</td>
<td>-2.1691 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERCC2</td>
<td>19</td>
<td>50547-50566</td>
<td>0.0543 (0.0%)</td>
<td>-0.4355 (5.6%)</td>
<td>-0.5161 (0.0%)</td>
<td>0.0324 (0.0%)</td>
<td>-0.5733 (0.0%)</td>
<td>-2.8837 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATRN</td>
<td>20</td>
<td>3400-3580</td>
<td>0.0191 (0.0%)</td>
<td>1.2650 (6.9%)</td>
<td>-0.4662 (0.0%)</td>
<td>0.0367 (2.6%)</td>
<td>-0.9906 (7.8%)</td>
<td>-1.2372 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASIP</td>
<td>20</td>
<td>32312-32321</td>
<td>0.0781 (0.0%)</td>
<td>-2.2292 (0.0%)</td>
<td>-2.7555 (100.0%)</td>
<td>0.0983 (0.0%)</td>
<td>-2.8331 (0.0%)</td>
<td>-4.2391 (100.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDN3</td>
<td>20</td>
<td>57309-57334</td>
<td>0.0153 (1.4%)</td>
<td>-0.0043 (8.3%)</td>
<td>0.3102 (0.0%)</td>
<td>0.0305 (0.0%)</td>
<td>-0.9651 (8.6%)</td>
<td>-0.7886 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOX18</td>
<td>20</td>
<td>62150-62151</td>
<td>-0.0043 (0.0%)</td>
<td>2.6919 (0.0%)</td>
<td>-0.0450 (0.0%)</td>
<td>0.0146 (0.0%)</td>
<td>0.7989 (0.0%)</td>
<td>n/a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPS4</td>
<td>22</td>
<td>25177-25210</td>
<td>0.0652 (3.2%)</td>
<td>-0.6695 (1.6%)</td>
<td>-0.2946 (0.0%)</td>
<td>0.0821 (3.6%)</td>
<td>-0.7824 (3.6%)</td>
<td>-1.9946 (0.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOX10</td>
<td>22</td>
<td>36698-36713</td>
<td>-0.0237 (0.0%)</td>
<td>-0.4479 (13.3%)</td>
<td>1.3746 (0.0%)</td>
<td>0.0182 (0.0%)</td>
<td>0.0137 (0.0%)</td>
<td>-1.0609 (0.0%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Identification of Selection-Nominated Candidate Genes

Based on these findings, 14 genes, *ADAM17, AP3D1, ASIP, ATRN, EGFR, HPS1, KIT, MYO5A, OPRM1, PAX3, PLDN, POMC, RAB27A,* and *SILV,* showed evidence of selection unique to the Indigenous American populations. These 14 genes are referred to as selection-nominated candidate genes in reference to the fact that they are known both to be involved in biochemical pathways influencing skin pigmentation and to show evidence of selection in Indigenous Americans. Figure 2-6 depicts the LsBl, *lnRH,* and Tajima’s D for these 14 genes. This figure also contains plots of linkage disequilibrium showing the haplotype structures of these genes in the Indigenous American and East Asian populations. These figures contain only the genes showing evidence for selection, graphical representations of the results for the remaining genes can be found in Appendix B.
(a) ADAM17

- ADAM17 LiRFH HGDP
- ADAM17 LSL/H HGDP
- ADAM17 Tajma's D HGDP
- ADAM17 Tajma's D Difference HGDP

- ADAM17 LiRFH Mega
- ADAM17 LSL/H Mega
- ADAM17 Tajma's D Mega
- ADAM17 Tajma's D Difference Mega

- D' HGDP America
- D' HGDP East Asia
- D' Mega America
- D' Mega East Asia
(b) $AP3D1$
(c) ASIP

ASIP LDRH HGD

ASIP LSLH HGD

ASIP Yploc's D HGD

ASIP Yploc's D Difference HGD

ASIP LDRH Mega

ASIP LSLH Mega

ASIP Yploc's D Mega

ASIP Yploc's D Difference Mega

D' HGD America

D' HGD East Asia

D' Mega America

D' Mega East Asia
(d) \textit{ATRN}
(e) **EGFR**

![Graphs and Heatmaps](image-url)

D' HGDP America  
D' HGDP East Asia  
D' Mega America  
D' Mega East Asia
(f) *HPS1*
(g) KIT

D' HGDP America  D' HGDP East Asia  D' Mega America  D' Mega East Asia
(h) MYO5A
(i) OPRM1
(j) P4X3
(k) PLDN

D' HGDP America

D' HGDP East Asia

D' Mega America

D' Mega East Asia
(1) POMC

![Graphs showing POMC](image)

<table>
<thead>
<tr>
<th>D' HGDP America</th>
<th>D' HGDP East Asia</th>
<th>D' Mega America</th>
<th>D' Mega East Asia</th>
</tr>
</thead>
</table>
(m) RAB27A
In addition to the evidence indicating selection in this research, there exists for each of these genes an immense body of information from mouse coat color research, statistical genetics and population studies. Thus, the discussion of each candidate gene will be limited to a brief review of the functional importance and a summary of the evidence of selection. Figure 1-4 illustrates the related roles of many of these selection-nominated candidate genes in the production of melanin. For all of the reasons discussed in the section on the necessity of employing multiple tests, it is unreasonable to expect that the results will agree in all cases. The differences in time-depth, ability to identify selection on standing variation, etc. will result in discrepancies which will be discussed for each gene.

ADAM17 is involved in melanocyte development and maturation and encodes the major phorbol-ester-stimulated sheddase which modulates the release of the membrane-bound enzyme kit ligand which, once made soluble by release from the membrane, is essential in melanogenesis [4, 71]. This gene also shows strong evidence of selection in East Asian populations [30, 34] as well as in the Indigenous American populations investigated here. ADAM17 shows evidence of selection in both samples at lnIRH and Tajima’s D. OPRM1 shows strong evidence of selection in both samples for LSBL with a particular enrichment of the downstream end of the gene. Tajima’s D difference values in this same region of the gene also indicate selection. Polymorphisms in the gene have been associated with variation in the relative amounts of eumelanin and pheomelanin produced [72]. SILV is a homologue of a mouse coat color gene and has previously been associated with iris color in humans [73]. In vitro studies indicate that it catalyzes an intermediate step in the conversion of dopaquinone into eumelanin [74]. Poorly
covered with SNPs, *SILV* shows large Tajima’s D difference values in the Mega data. *PLDN* is one of a family of enzymes involved in the biogenesis of various organelles including melanosomes as well as lysosomes, and granules. [75] The mouse model indicates that it is involved in membrane fusion [76] and that mutations in the gene result in defects in melanin production. *PLDN* is noteworthy for the haplotype structure in the Indigenous American populations. For both samples, there appears to be an increase in the size of a haplotypes in the downstream region of the gene. Additionally, there are significant results for LSBL and lnRH results for both samples and the Tajima’s D difference values are negative, though not significant, and show an increase over the Tajima’s D values. *MYO5A* and *RAB27A* are known to interact to form a complex composed of a GTP-ase encoded by *RAB27A*, a vesicle motor encoded by *MYO5A*, and melanophilin. This complex is essential for the transport of melanosomes [77]. Additionally, *MYO5A* has been previously associated with iris color in Europeans [73]. The evidence for selection in *MYO5A* is stronger in the Mega sample than in the HGDP, with significant Tajima’s D difference results in addition to the significant lnRH values found in both samples. *RAB27A* shows some evidence of selection in the upstream region of the gene as indicated in both samples for LSBL with support from the HGDP lnRH results and the appearance of a haplotypes block in the Indigneous Americans for the HGDP sample. Curiously, the Tajima’s D and Tajima’s D difference results are in opposite directions for the HGDP and Mega samples with the HGDP results showing significantly negative results and the Mega sample showing positive results. This discrepancy could be due to allele frequency differences among the two populations or to the different densities of markers in the two samples. The Mega data appears to have denser marker coverage in the region of the gene where the discrepancy occurs.

Previous reports indicate that *ASIP* shows evidence of selection in both East Asians and Europeans, which suggests a selective event prior to the divergence of these populations [34]. The lnRH and LSBL values for this gene are in the right direction to indicate selection but are not
significant. However, ASIP has some of the largest Tajima’s D difference values anywhere in the genome. ASIP encodes the agouti signaling protein which, in the mouse model, acts as an inverse agonist in the binding of the α-melanocyte stimulating hormone (α-MSH) to Mc1r which regulates the relative production of eumelanin and pheomelanin [78]. ATRN also shows evidence of selection in East Asians [34] but also is distinguished by significant results on lnRH, LSBL, and Tajima’s. ATRN encodes a type-I transmembrane protein which acts as an accessory receptor for the agouti protein [79]. POMC encodes the pituitary hormone pro-opiomelanocortin which is the pre-cursor to α-MSH [80] and has been identified as a gene potentially involved with iris color variation in Europeans [73]. The gene shows a substantial increase in the size of a haplotypes block in the upstream region in Indigenous Americans compared East Asians for the HGDP sample. This is supported with LSBL results in both samples indicating evolution in the Indigenous Americans since the split with the East Asian populations.

AP3D1 is a gene associated with the biogenesis and trafficking of pigmentation molecules [81]. The significant values for lnRH and LSBL in the HGDP samples, although interestingly not the Mega data are joined by the strongly negative, although just shy of significant values for Tajima’s D difference. HPS1 interacts with the AP3 complex associated with the membranes of organelles [82]. This gene shows significantly negative values of lnRH at the downstream end of the gene and a haplotypes block and positive LSBL values at the upstream end. EGFR is an epidermal growth factor receptor implicated in a wide variety of functions in skin cells including cell proliferation, differentiation, and mobility and is also involved in the developmental process [83]. EGFR shows a substantial increase in haplotypes block size on both the Mega and HGDP samples as well as isolated significant results in the LSBL and lnRH. The Tajima’s D difference values for this gene are particular striking in the HGDP sample, making it surprising that the Mega Data shows positive values over some of the same region. PAX3 interacts with SOX10 to activate MITF which is integrally involved in melanocyte proliferation.
Mutations in *PAX3* have been implicated in pigmentation disorders. [84] This gene shows strong evidence of selection with significant values in all three tests on both samples. *KIT* encodes a tyrosine kinase receptor which, with its ligand *KITLG*, is involved in hematopoiesis and gametogenesis in addition to melanogenesis [85]. *KIT* has a possible enlargement of a haplotype in the Indigenous Americans as well as isolated significant results in most of the tests.

**Selection on Pigmentation Genes as a Class**

This data can also illuminate a number of questions about pigmentation genes as a class. It has previously been shown that regions surround pigmentation genes tend to have higher $F_{st}$ values in pairwise comparisons between Europeans and Africans and between Europeans and East Asians than randomly selected regions of the genome. [36] A similar result was found in this analysis. The mean values for each test were compared for all SNPs in the pigmentation gene regions and all other SNPs. Welch’s $t$-tests were performed because this statistic is robust to unequal variance [86]. With the exception of the lnRH statistic for the Mega sample, the values for SNPs in the regions surrounding pigmentation genes were on average more extreme than those for all other SNPs, as can be seen in Table 2-4. The non-pigmentation gene SNPs include both SNPs in and around other genes which may be under selection of some kind and a great number of SNPs that are presumably evolving neutrally. This suggests that pigmentation SNPs as a class are more likely than randomly chosen SNPs to be undergoing selection in Indigenous American populations. Future research may focus on the estimation of the precise strength of selection on these genes.
Table 2-4. Comparison of Selection in Pigmentation and Non-Pigmentation Genes.

<table>
<thead>
<tr>
<th>Test</th>
<th>Sample</th>
<th>Non-Pigmentation Genes</th>
<th>Pigmentation Genes</th>
<th>t-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lnRH</td>
<td>HGDP</td>
<td>-0.6643 (2.1739)</td>
<td>-0.8221 (2.4063)</td>
<td>0.0012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mega</td>
<td>-0.0966 (2.4359)</td>
<td>-0.1710 (2.5331)</td>
<td>0.0957</td>
<td></td>
</tr>
<tr>
<td>LSBL</td>
<td>HGDP</td>
<td>0.0572 (0.0931)</td>
<td>0.0704 (0.1100)</td>
<td>1.33E-08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mega</td>
<td>0.0444 (0.0866)</td>
<td>0.0580 (0.1056)</td>
<td>3.21E-11</td>
<td></td>
</tr>
<tr>
<td>Tajima's D Difference</td>
<td>HGDP</td>
<td>1.5831 (0.9994)</td>
<td>-1.7479 (1.1167)</td>
<td>0.0011</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mega</td>
<td>0.0014 (0.9988)</td>
<td>-0.2848 (1.1981)</td>
<td>1.02E-07</td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

This research has demonstrated for the first time evidence of selection at multiple skin pigmentation genes in New World populations. Although there is substantial statistical evidence for selection in these skin pigmentation genes, this evidence alone should not be treated as a conclusive result. To identify genes contributing to the observed differences in skin pigmentation among populations, polymorphisms in these genes must show an association with variation in the phenotype. Nevertheless, it is a compelling first step in identifying genes that responded to the new environmental pressures as humans first colonized the Americas.
References


78. Rieder, S., et al., *Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus)*. Mammalian Genome, 2001. **12**(6): p. 450-455.


Chapter 3

Admixture, Ethnicity, and Skin Pigmentation in Southwestern Colombia

Before selection-nominated candidate genes can be assessed via admixture linkage analysis, the biogeographical ancestry (BGA) of the populations involved must be characterized. Ancestry Informative Markers (AIMs) is a broad term that encompasses any genetic marker that shows substantial allele frequency differences between populations and can therefore be used to identify the likely BGA of an individual. In other words, AIMs can be used to estimate the population or populations to which an individual or their ancestors most likely belonged. AIMs are of scientific value for a number of reasons including in forensics to estimate the likely ancestry of a suspect or victim, to reconstruct admixture and prehistoric migration patterns, and to control for confounding stratification in genetic association or epidemiological studies. All of these uses have been discussed thoroughly in other publications.

The use of these markers to access the recent history of admixture in populations is the focus of this work. Understanding recent admixture is important for both anthropological and genetic reasons. The use of outmoded racial and ethnic designations treats groups who are heterogeneous at both the phenotypic and genotypic level as homogenous populations. The deleterious effects of race-based stereotypes are widely known and a better understanding of the overlapping ancestries of most of human populations can and should be used to undercut these assumptions. Of additional concern is the use of these same homogenizing labels in so-called “race-based” medicine. Admixed populations such as African Americans and Hispanics contain broad ranges of genetic ancestries which can be related to particular phenotypes. However, treating these populations as single groups obfuscates these associations and/or creates false positives, minimizing the ability to detect genes that may underlie these phenotypes [1]. As will be further discussed in the next chapter, establishing individual admixture estimates is the critical
first step in admixture linkage analysis which can be used to identify genes related to population-specific differences in phenotypes.

Limitations and Merits of Different Types of AIMs

There are two main types of AIMs: uniparental markers, which identify ancestry along either the maternal line (via mitochondrial DNA) or paternal line (via Y-chromosomal markers) and biparental (autosomal and X-chromosomal) markers which provide an average across all ancestors. Generally, the uniparental markers have a greater degree of specificity while the biparental markers provide a more inclusive view of ancestry. These differences are due to the different inheritance patterns among these markers as illustrated in Figure 3-1.
Figure 3-1. Inheritance of uniparental and biparental markers. The male (square) and female (circle) offspring at the bottom of this pedigree have both inherited their mtDNA (shown as a ring) from their maternal grandmother while the male offspring has inherited a Y-chromosome (small linear chromosome) from his paternal grandfather. Y-chromosomes in the maternal lineage and mtDNA in the paternal lineage are lost. The pairs of autosomal chromosomes (represented here by a single pair) are inherited from each parent and include portions of chromosomes from all grandparents due to recombination during meiosis. Figure modified from [2].

**Uniparental Markers**

Mitochondrial DNA (mtDNA) is a ring of DNA found in the mitochondria distinct from the nuclear DNA chromosomes. MtDNA is a robust genetic marker favored in molecular anthropology communities because there are hundreds to thousands of identical copies per cell as opposed to just two copies of nuclear DNA. This enables viable DNA to be extracted from highly degraded samples including ancient DNA as well as relatively modern, but poorly preserved samples. Children inherit mitochondria exclusively from their mothers as mitochondria are abundant in the cytoplasm of eggs but very rare in sperm cells. As a result, mtDNA is exclusively informative of the maternal lineage. Relative to nuclear DNA, mtDNA has a high mutation rate and, since mtDNA is not subject to recombination, mutation is the only source of
new variation in the sequence. Added to this the smaller effective population size of mtDNA, which increases the influence of drift, makes mtDNA ideal for the study of population structure.

[3]

AIMs on the Y chromosome are predominantly if not exclusively located in the non-recombining (NRY) portions of the chromosome, as opposed to the pseudo-autosomal region (PAR) which can align and cross over with the X chromosome. Because NRY markers have no homologues on the X chromosome, they can only be analyzed in males. Like mtDNA, NRY markers are only influenced by mutation and have a smaller effective population size than nuclear DNA so they are a reliable and useful marker of inheritance in the paternal lineage. [4]

One major advantage of uniparental markers for inferring ancestry is the ability to identify a likely geographic region of ancestry based on shared haplotypes with the modern inhabitants. A haplotype is a set of alleles that are inherited together. Because mtDNA and NRY do not undergo recombination during meiosis, the entire molecule can be considered to have a single haplotype – unlike nuclear DNA which is comprised of many haplotypes. Similar haplotypes can then be combined into haplogroups which may correspond to particular regions of the globe. With appropriate adjustments for recent migrations and admixture, the phylogenetic trees for both mtDNA and NRY show the most diversity in Africa with all non-African haplotypes descending from African haplotypes as would be expected for the evolutionary history of our species [5]. When determining ancestry, it is important to consider that it is difficult to estimate time depth for uniparental ancestry markers. A man with a Y haplotype belonging to the most common haplogroup in Wales could have a Welsh father, Welsh paternal grandfather, or a Welsh paternal ancestor twenty generations back. Human migrations and genetic drift during that intervening time will also complicate lineage matching.

From the standpoint of estimating individual ancestry, the predominant problem with the use of uniparental markers alone is that the vast majority of information about an individual’s
ancestry is left unmeasured. Consider Figure 3-1 where uniparental markers can provide the male offspring with information about his paternal grandfather and maternal grandmother or 50% of his ancestors two generations back. For the female offspring, for whom NRY markers cannot be genotyped, only 25% of her ancestors two generations back can be considered. Estimating ancestry using uniparental markers even as few as five generations (or approximately 100 years) in the past means that the male offspring has no information about 94% (30 of 32) of his ancestors and the female offspring had no information about 97% (31 of 32) or her ancestors.

**Autosomal Markers**

Autosomal markers overcome the limited perspective on ancestry afforded by uniparental markers but have limitations of their own. Most importantly, as biparental markers, autosomal AIMs produce are informative for all ancestors, not simply maternal or paternal lines [4]. However, due to recombination, the time depth over which these AIMs are informative is shorter than for uniparental markers. AIMs are relatively rare in the genome because the vast majority of genetic variation is found within populations not between them. This is illustrated by genome-wide mean Fst value between pairs of globally distributed populations which is only 0.148 [6]. In particular, the more closely related two populations are, the more difficult it becomes to identify a sufficient number of AIMs to differentiate between the populations. The clinal distribution of variation in our species and the relative recent shared ancestry of all humans exacerbates these difficulties. However, large-scale SNP genotyping over the past five years has rapidly increased the number of available autosomal AIMs and increased the overall understanding of the substructure found in many populations.

This increase in known AIMs allows for fine-scale calculation of proportional BGA across the genome as well as for individual loci. Proportional and locus-specific estimates of
ancestry throughout the chromosome are essential for accurately controlling for ancestry when performing association studies [7]. Uniparental markers simply cannot be used for these types of analyses because those markers can only measure proportional ancestry at the population level not the individual level. Many autosomal AIMs panels have been developed for measuring individual BGA in different populations, with more AIMs needed for use in populations with longer histories of admixture, more ancestral populations, or more recent parental population divergence times and therefore smaller $F_{st}$ values. Panels have been identified for use in Hispanic populations [8-10], African-American populations [11-13], and within Europe [14]. The use of admixture estimates in linkage and association studies will be discussed in detail in Chapter 4.

**Admixture Studies in South America**

Positioned on the only land route into South America, modern-day Colombia was almost certainly the first part of South America to be colonized by humans. The earliest archaeological sites in Colombia, Tibito and El Abra, place the first human settlements in the region no later than 12,000 years ago [15]. Spaniards arrived as early as 1500, establishing in what would later be the northern coast of Colombia the first non-Indigenous American settlements in the New World. West African slaves were brought to Colombia soon thereafter, predominantly to work in the gold mines and cacao plantations. [16] By the end of the 18th century, more than 10% of all Spanish slaves in the New World, more than 7,000 individuals in total, toiled in the Colombian gold mines alone [17]. Since that time, the tripartite ancestry of the Colombian people has been influenced at different times by sentiments and policies that encouraged or discouraged mating among Indigenous populations, Spaniards, West African slaves and later free Africans. [18] Additionally, Colombia is a large country with tremendously varying terrain including large
mountain ranges that may have led to the relative isolation of some populations. The net result of this shared history is a complex distribution of ancestries and ethnicities across Colombia.

Sex-Biased Gene Flow

From historical, demographic, and biological perspectives, it is important to understand the admixture patterns in South America. One means of doing this is to investigate the distributions of AIMs in the populations to estimate admixture proportions. In theory, uniparental and biparental markers should provide the same information at the population level. For example, in an admixed population where 50% of the ancestors of the modern population came from Western Europe and 50% came from Sub-Saharan West Africa, one would expect that approximately 50% of mitochondrial, 50% of Y-chromosomal, and 50% of autosomal markers would be traced to each of these regions. However, in practice, these methods of estimating ancestry are not equivalent. A number of studies discussed in the following sections have found that the ancestry of the mitochondrial DNA and Y-chromosomes differs substantially within populations, a common pattern in many admixed African-American and Hispanic populations in the Americas [13, 19-23]. This evidence is consistent with differential contributions to the gene pool between males and females. In many cases, European Y-chromosomes are over-represented in these populations which is generally interpreted as a history of European males mating with African or Indigenous American females at a greater rate than African or Indigenous American males mated with European females. An alternate interpretation would be that children of European mothers and African or Indigenous American fathers would be less likely to be considered African-American or Hispanic but the history of racial classification by hypodescent in the United States and much of the rest of the Americas does not support this conclusion. A final possibility is that, while mtDNA and NRY markers are generally considered to be neutral,
selection in other regions of the mtDNA and Y chromosome may increase or decrease the
frequency of the markers since they are located in non-recombining regions of the genome [24,
25].

Sex-biased gene flow is not an exclusively recent phenomenon nor is it restricted to the
New World. Differences in ancestry estimates between NRY and mtDNA markers have been
identified in Tibeto-Burman populations in China [26], between Bantu farmers and indigenous
hunter-gatherers during the Bantu expansion in Africa [27], and for Scandinavian explorers in
Greenland [28] among other examples. Once these mating asymmetries occur, they persist in the
sex-specific markers due to lack of recombination and so can be detected many generations later.

**Admixture in Colombia**

Biogeographical ancestry (BGA) has been studied in a number of regions of Colombia as
part of efforts to understand recent ancestry as well as recapitulate the peopling of South
America. The most striking result of these many studies is the regional variability of estimates
among the same ethnic groups and the variation among ethnic groups in the same and different
regions. According to the 2005 Census, the Colombian population is 58% Mestizo
(European/Indigenous American), 20% White (European), 14% Mulatto (European/African), 4%
Afro-Colombian, 3% Zambo (African/Indigenous American), and less than 1% Indigenous
American [29]. These are nation-wide averages and there is tremendous variability in
distributions of ethnic groups and ancestries among the various departments that comprise the
country of Colombia. The geographic and political boundaries of Colombia can be seen in Figure
3-2. These self-reported ethnicities also indicate a clear acknowledgement of admixture in the
country with 75% of individuals reporting an admixed ethnic group (Mestizo, Mulatto, or
Zambo).
Salas et al. [30] report that in three admixed populations collected predominantly in the Western coastal departments of Cauca and Valle del Cauca, 97% of self-identified Mestizos carry an Indigenous American mtDNA haplotype while 81% of Mulattos and 76% of Afro-Colombians have African mtDNA haplotypes with the remaining haplotypes Indigenous American. These mtDNA results are similar to those found for Afro-Colombians from other Pacific and Caribbean coastal populations and Mestizo individuals from Bogota studied by Rodas et al. [31]. In contrast, other studies have found substantially higher levels of European ancestry (greater than
60%) in Mestizo populations in the mountainous northern department of Antioquia based on autosomal markers [32].

Mesa et al. [33] compared mtDNA, NRY, and autosomal markers in five populations living in resguardos (native territories) in the departments of Amazonas, Guarjira, Putumayo, Córdoba, and Antioquia. Autosomal averages for Indigenous American ancestry ranged from 69% to 100% among these populations but the more interesting results come from the uniparental markers. The Indigenous American ancestry is uniformly high for the mtDNA with 95% in the most admixed group. However, in some populations, the proportion of Indigenous American Y haplotypes is less than half with both African and European haplotypes occurring frequently. This is clear evidence of a sex-biased gene flow with African and European males having mating with Indigenous females at high levels. This result is particularly interesting because the subjects all self-identified as members of specific indigenous groups and speak a tribal language. A number of studies have focused on indigenous Colombian populations in an attempt to decipher the patterns of human colonization of the Americas [34-36]. While the study of mtDNA in these populations should yield reliable results, this evidence of admixture must be considered in the interpretation of any other marker system. A bias towards European male chromosomes has been found in other ethnic groups in Colombia [37, 38]. One of the most extreme examples of sex-biased gene flow can be seen in the relatively isolated Antioquia population which occupies a mountainous region of northwestern Colombia. Approximately 94% of the Y haplotypes in this population are European in origin while in contrast 90% of the mtDNA haplotypes are Indigenous American [39].

The broadest study of ancestry, ethnicity, and sex-bias gene flow to date is awaiting publication by Rojas et al. [40]. In this study, 1,700 individuals were sampled from 24 predominantly Mestizo and Indigenous American populations distributed around Colombia. These researchers identified a number of populations that are geographically near one another but
quite different in their autosomal ancestry estimates. This local variation in BGA estimates is not
unique to Colombia but certainly places it among the most extreme examples of broad variability.
In nearly every population, there were significantly more Indigenous American mtDNA
haplotypes than Y haplotypes with the majority of Y haplotypes coming from European and to a
lesser extent African ancestors. The results of this study are not surprising considering the
previous work that has been done, but the breadth of the study shows that these differences are
unlikely to be sampling bias or due to systematic differences among the study protocols.

Admixture in Neighboring Countries

Admixture patterns have also been studied in detail in many other South American
countries increasing our understanding of both historic and prehistoric population movements.
Brazil is likely the most studied country in South America and, like Colombia, the modern
population is predominantly composed of admixed descendents of Indigenous Americans,
European colonists, and West African slaves. The majority of the Portuguese contribution to the
gene pool in the two centuries following their arrival in 1500 was male and this type of
directional mating was actively encouraged by the Portuguese government as a means of
population growth and colonial control [41]. According to historical records, 3.5 million African
slaves were brought to Brazil between 1551 and 1850 and nearly 500,000 Portuguese, voluntarily,
immigrated to Brazil as well. In the past century and a half, the immigration has remained
predominantly western and southern European. [41] As a result, virtually every study of genetic
or genomic ancestry in Brazil has shown admixture, but the amount varies between different
markers, different regions, and different reference populations. Alves-Silva et al. found
approximately equal proportions of Indigenous American, European, and West African
haplotypes in a sample collected from four of the five main geographic regions in Brazil [41].
Although this population primarily self-identified as “white” (branco), this evidence strongly supports a female ancestry contribution from Indigenous and African women. This evidence is in clear contrast to the Y-chromosomal data which suggests that self-identified white males from across Brazil have almost exclusively European-derived Y-chromosomes [42]. As expected in such a large country, the sub-continental ancestry of these Y haplotypes is diverse. In southern Brazil, many of the white males have Y-chromosomal haplotypes most similar to Italians while in the Amazon there is a large proportion haplotypes similar to those found in Moroccan Jewish men. Mirroring the 17th century invasion of northeast Brazil by the Dutch, males there have a higher proportion of Dutch-derived haplotypes. In that sample, only 2.5% of the males had African Y-chromosome haplotypes and none were Indigenous American. [42] When the same analyses are performed among self-described “black” (preto) individuals, the proportion of African Y-haplotypes is larger but a large proportion of European Y-haplotypes is still found, particularly in urban populations [43]. In a recent study of individuals from the Rio Grande do Sul, the southern-most state in Brazil, 79% of mtDNA sequences had an African origin compared to only 44% of the Y-chromosomes. The Indigenous component was small but present in both the mtDNA and Y-chromosomes. [44] Blending what is known from historical records with data from AIMs provides a clearer picture of the history of this diverse country than either written records or genetics could produce on its own.

Similarly, studies in two of Colombia’s other neighbors, Venezuela and Ecuador, have also shown evidence of directional mating among their highly admixed populations. In Ecuador, Mestizos are by far the largest ethnic group comprising 60% of all Ecuadorians. Individuals identifying as indigenous – mostly Kichwa – are the next most common and live predominantly in the highlands and plains regions. Afro-Ecuadorians comprise the remainder of the population and reside largely in rural areas in a particular portion of the Andes and in coastal regions. Among these three distinct groups, there are mean differences in ancestry but there is also
substantial overlap. González-Andrade et al [45] analyzed autosomal and NRY STRs to estimate the ancestry of these three ethnic groups and found that in the Mestizo population, Y haplotypes were approximately 70% European, 28% Indigenous, and 2% African in origin where the relative contributions in the autosomal markers averaged 19%, 73%, and 8%, respectively. The sexual asymmetry in mating, likely at the establishment of the Mestizo group, is abundantly clear. The prevalence of European Y haplogroups extends to the Indigenous populations as well where 10% of males who self-identified as completely Kichwa were found to have European Y chromosomes. Among the Afro-Ecuadorans, who maintain a great deal of phenotypic and cultural distinction from the other ethnic groups, Y haplogroups are 44% African, 31% European, and 15% Native American. This greater level of African male ancestry is to be expected but it is worth noting that the proportion of African Y chromosomes is still lower, and of European Y chromosomes still higher, than the autosomal estimates of 56% African, 16% European, and 28% Amerindian.

A study by Martínez et al. in Caracas, Venezuela [46], illustrates another important dynamic in assessing admixture within and among ethnic groups in South America and elsewhere – socioeconomic status. In this urban sample, individuals of higher socioeconomic standing (SES) are predominantly European in ancestry while individuals of lower standing have a substantially larger proportion of both Indigenous American and African ancestry. In both the higher and lower socioeconomic classes, Indigenous American mtDNA haplotypes are extremely common approximately 40% and more than 70%, respectively. When Y haplotypes are considered, more than 90% of males of high SES have a European Y-chromosome, compared to 84% of low-SES males. The discrepancies between the male and female uniparental markers echoes the early days of colonialism when few European females were present in Colombia as well as the civil wars and the war of independence which resulted in the deaths of large numbers of indigenous men in the late 19th century. The difference in ancestry proportions between groups
of higher and lower socioeconomic status is illustrative of circumstances found in many populations where individuals with European ancestry maintain greater access to power and wealth even after colonial rule has ended.

The above studies do not comprise a complete picture of admixture in South America, but are representative of the general findings. The use of exclusively uniparental or biparental markers is insufficient to understand the history of admixture in these countries. Throughout the region, researchers have repeatedly found evidence of European-biased male contributions to Indigenous and African populations. Furthermore, these ancestry differences are in some cases tied to overarching differences in social status. To better understand how the admixture in western Colombia compares to these neighboring counties, analysis of Y-chromosomal and autosomal AIMs was undertaken.

**Samples and Methods**

**Description of Samples**

The samples used in this analysis were collected by Marc Bauchet (Max Planck Institute for Evolutionary Anthropology) in Popayán, Cauca, Colombia as part of a broader study overseen by Mark Stoneking. The majority of individuals were students at the University of Cauca with the remainder residents of the local area. A questionnaire was used to identify languages spoken and self-identified ethnicity, parental, and grandparental ethnicity, and demographic information including place and date of birth. To supplement the standard form, interviews were conducted that included the language spoken by informant and family members, geographical origins, customs or traditions, marriage patterns, and reproductive history. During this interview, pedigrees were generated linking many of the participants. Saliva samples were collected for
DNA analysis and skin pigmentation was measured using the DSM II ColorMeter (Cortex Technology, Hadsund, Denmark) with both Melanin (M) Index and Luminance (L*) recorded. Digital photographs were taken of the frontal face and profile as well as close ups of the eyes for iris color analysis (Sony, Tokyo, Japan). Finally, height and weight were measured, hair samples were collected, and dental casts were taken for further analysis. All sample collection was done with informed individual consent as well as IRB approvals at the Max Planck Institute and Pennsylvania State University.

Three-hundred and eighty-two individuals were sampled (201 females, 181 males), however many individuals were first-degree relatives based on pedigrees obtained during sample collection. Once related individuals were removed, the final data set includes 139 females and 127 males. Individuals in the sample predominantly self-describe as Mestizo (75% of individuals) with 11% identifying as Campesino, 9% Indigena, and 6% Afro-Colombian. Ethnographic and political science research suggests that the terms Indigena and Campesino are strongly related to each other with “campesino,” meaning peasant, in declining use with a surge of pro-Indigenous political movements in the late 1980s and early 1990s [47]. The term campesino tends to be used more frequently among the older generation and in rural areas (personal communication, Marc Bauchet) Based on these reports and findings of genetic similarities, these groups were combined for some analyses to increase sample size and statistical power.

**Y-Chromosome Genotyping**

Each male in the sample was genotyped at the Max Plank Institute for Evolutionary Anthropology in Leipzig, Germany for panel of 24 SNPs on the non-recombinant portion of the Y-chromosome (NRY SNPs). These SNPs (SRY, M207, M9, M106, M175, Tat(M46), M172,
M96, 12f2, M91, M145, M174, M69, M170, M45, M213, M52, M168, M269, M20, M214, M124, M201, and MEH2) represent branching nodes in the phylogenetic tree of human Y chromosomes. With reference to established gene trees, these SNPs can be used to identify each man’s Y haplotype. The phylogenetic tree in Figure 3-3 illustrates how these SNPs distinguish between the haplogroups. The Y haplogroups shown here can be further subdivided with additional SNPs to provide greater resolution.

Figure 3-3. Diagnostic SNPs for Y haplogroups in sample. Every Y haplogroup is identified by one or more diagnostic mutations that are unique to the haplogroup. At each branching point, the diagnostic SNP is indicated. Image from genebase.com.

**Autosomal AIMs Genotyping**

All individuals in the sample were genotyped for 106 autosomal AIMs [48] at the University of Minnesota BioMedical Genomics Center. The SNPs on this AIMs panel were selected for maximal allele frequency differences between West African, European, and Indigenous American parental populations and even spacing across the 22 autosomal chromosomes. Allele frequencies for AIMs can be found in Appendix III. Genotyping was
performed using the Sequenom iPLEX Gold assay based on PCR followed by single base extension. The multiplex is them read by the MassARRAY platform which is a mass spectrometry system can discriminate between the base extension products for up to 40 SNPs per well. Individuals with less than 100 successfully genotyped AIMs were not included in the analysis as previous simulations in other Latin American populations with tripartite ancestry support this cutoff as the minimum necessary to produce a reliable estimate of genome-wide ancestry using this panel of AIMs. With at least 100 AIMs, there is a greater than 0.9 correlation coefficient with known ancestry proportions as well as controlled type I error at the 5% level.

[49] Point estimates of genomic ancestry were calculated using maximum likelihood estimation (MLE) implemented with a perl script written by Vibhor Sonpar. MLE produces ancestry point estimates and confidence intervals for each individual by comparing observed alleles to frequency distributions in the parental populations. MLE will be discussed in greater detail in the next chapter along with additional methods for estimating ancestry.

**Results and Discussion**

**Y-Chromosomal Ancestry Estimates**

Based on the 22 NRY SNPs, 123 unrelated males were found to have one eleven distinct Y haplogroups – B*, E, Q, K*, G, I, J*, J2, L, R1a, and R1b. The distribution of haplotypes by ethnicity can be seen in Figure 3-4, in which the Campesino and Indigena individuals have been grouped together in the pie chart. Although it is difficult to accurately perform a statistical test on the similarity between these two groups due to small sample sizes, the distribution of Y-chromosomes among these two groups appears more similar (Q is the most common followed by R1b and E) to one another than either is to Afro-Colombians or Mestizos.
Figure 3-4. Distribution of Y haplogroups by ethnicity. The eleven haplogroups found in the Colombian sample show differential distributions across the three ethnic groups. The Mestizo group is the largest and contains the most diversity.

Based on their modern distributions and phylogenetic reconstructions from sequence data, each of these haplogroups has a previously inferred geographic origin [50]. At the population level, the Colombians have a large proportion (40%) of R1b haplotypes, which is not surprising considering R1b is among the most common haplogroups in Western Europe. The next most common Y haplogroups are Q (24%) and E (18%) which are Indigenous American and
African in origin respectively. The worldwide distributions of these three haplogroups can be seen in Figure 3-5.

Figure 3-5. World-wide distribution among indigenous groups of most common Colombian haplotypes. The three most common haplogroups – R, Q, and E – found in this sample are representative of the tripartite ancestry of the population. Figure modified from [51].

The distribution of haplogroup R1b in Europe increases clinally from East to West, reaching a peak in the British Isles where the prevalence ranges from 65-83% [52]. Based on phylogenetic reconstructions of variation within the haplogroup, the origins are most likely in the Near East approximately 18,500 (CI: 12,500-25,700) years before present [50]. A number of researchers have associated the expansion of the R1b1b2 haplogroup and other European haplogroups with the expansion of farming during the Neolithic [53]. An alternate explanation attributes the modern distribution of the R1b haplogroup in Europe to cultural diffusion out of the Iberian peninsula, where R1b is also at high frequencies, following the last glacial maximum [54]. The highest frequency of the R1 haplogroup outside of Europe is among the Hausa of Sudan. This is in keeping with previous evidence of this population’s back-migration from Western Asia into northeastern Africa. [55]

The Q haplogroup is one of the core haplogroups of the Americas with frequencies as high as 77% across indigenous populations in the northern portions of South America [56]. The recently sequenced 4,000 year old Greenland Paleo-Eskimo also had a Q haplotype testifying to the broad distribution of this haplogroup throughout the Americas [57]. Although modern distributions show Q predominantly in the New World, the origins of the haplogroup was likely a
mutation in a man with a P haplotype living in Central Asia or Siberia approximately 15-18,000 years ago [58]. Among modern Old World populations, the highest frequency of Q haplotypes is found in Siberia which coincides with Indigenous American origins and the most likely path of entry into the New World [59].

Based on STR-dating, Haplogroup E is an old clade likely originating in East Africa approximately 52,500 (CI:44,600–58,900) years ago. This old date is supported by the fact that E is the most diverse of all haplogroups having accumulated many mutations as it spread throughout Africa. Although it is at highest frequency in West Africa in modern populations, there are substantial frequencies outside of Africa in the Middle East and Southeastern Europe [50]. The particularly high frequencies in coastal West Africa from whence a large proportion of African slaves were kidnapped by the Spanish is congruent with the frequency of the E haplogroup in Colombia.

Based on the known distribution of these Y haplogroups in indigenous populations world-wide, the haplotypes present in the Colombian sample can be grouped into three major categories based on region of origin. Like E, B* shows highest frequencies in West Africa which suggests that Colombian men with these haplotypes likely descend along the paternal lineage from West African slaves brought to Colombia by the Spanish colonists. Men with a G, I, J, L, or R haplotype likely descend from a European male along their patriline and in the case of the many individuals with R haplotypes, this individual was likely Spanish given the history of the country and the high prevalence of the R haplogroup in the Iberian Peninsula. Males who belong to the Q haplogroup are likely descendent from Indigenous American males as this haplogroup is rare outside of the Americas with the exception of a region of northwestern Asia from whence there is little evidence of recent immigration to Colombia. The K haplogroup is unexpected in the Colombian population based on its origin in southwest Asia, although it is found at low frequencies though out much of the Old World [50]. Based on this information, the interaction
between self-identified ethnicity and patrilineal ancestry can be explored. As shown in Figure 3-6, the relationship between these two is complex.

<table>
<thead>
<tr>
<th>Y-Haplotype Origin</th>
<th>Ethnicity</th>
<th>Afro-Colombian</th>
<th>Campesino</th>
<th>Indigena</th>
<th>Mestizo</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>African</td>
<td></td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>European</td>
<td></td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>59</td>
<td>71</td>
</tr>
<tr>
<td>Indigenous American</td>
<td></td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>17</td>
<td>29</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>8</td>
<td>12</td>
<td>13</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>

![Figure 3-6. Y-chromosome origins by ethnicity. Y haplogroups were associated with likeliest region of patrilineal origin based on previous frequency and phylogenetic evidence. The degree of Y haplotype homogeneity among the ethnic groups is low.](image)

Although men who self-identify as Afro-Colombian are more likely to have an African Y haplotype than Mestizo, Campesino, or Indigena men, more than half of the African Y haplotypes identified in this analysis are found in Mestizos. Similarly, the distribution of European and Indigenous American haplotypes is roughly equivalent in the Campesino and Indigena individuals who self-identify as exclusively or predominantly of indigenous origins. Among the Mestizos, who acknowledge a large proportion of both European and Indigenous and of these three groups would be expected to have the highest levels of European ancestry, nearly two-thirds of males have a European haplotype. While it is clear that these ethnic groups are far from homogenous in terms of their ancestries, Fisher’s exact tests indicate that there are significant differences among the three ethnic groups in terms of Y haplotype ancestry. After a Bonferroni
correction, there are significant differences in the Y haplotypes distributions between Mestizos and Afro-Colombians ($p = 0.0111$), but not between Mestizos and Indigena/ Campesinos ($p = 0.0326$) or between Afr-Colombians and Indigena/Campesinos ($p = 0.0627$). The difference driving the significant result for the Mestizos-Afro-Colombian comparison is probably the large proportion of European haplotypes among the Mestizos and the large proportion of African haplotypes among the Afro-Colombians. In both ethnic groups, the proportion of Indigenous American Y haplotypes is low. This is particularly unexpected considering the known admixture between individuals of European and Indigenous American ancestry that produced the Mestizo population. Although large proportions of European ancestry are not unexpected among the Mestizos, it should be considered that high levels of European ancestry based on NRY data could be the result of European male-bias gene flow. To determine if these proportions of European ancestry are indicative of genome-wide trends or simply an increased proportion of European males mating with Indigenous or African females, autosomal AIMs were genotyped in both the males and females in the sample.

**Autosomal Ancestry Estimates**

As with the Y haplotype data, the MLE point estimates of BGA for individuals in the three major ethnic groups illustrate the heterogeneity found in this population and among the ethnic groups. As seen in Figure 3-7, there are substantial overlaps among the ethnic groups in terms of West African, European, and Indigenous American ancestry. Welch’s $t$-tests were calculated in R [60] to test for statistically significant differences in mean among the four groups. Welch’s $t$-test is an adaptation of the Student’s $t$-test for use in samples with unequal variances. Table 3-1 shows the results of these tests.
Table 3-1. Mean differences and $p$-values for Welch’s $t$-tests. Mean differences in proportional ancestry shown with $t$-test $p$-values in parentheses. Bold values indicate significant differences in mean after Bonferroni correction.

<table>
<thead>
<tr>
<th>Population Pair</th>
<th>African Ancestry</th>
<th>European Ancestry</th>
<th>Indigenous American Ancestry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afro-Colombian/Campesino</td>
<td>36% (9.37E-04)</td>
<td>4% (3.10E-01)</td>
<td>31% (5.50E-04)</td>
</tr>
<tr>
<td>Afro-Colombian/Indigena</td>
<td>44% (1.24E-04)</td>
<td>9% (5.93E-02)</td>
<td>35% (1.85E-04)</td>
</tr>
<tr>
<td>Afro-Colombian/Mestizo</td>
<td>39% (3.87E-04)</td>
<td>29% (1.77E-08)</td>
<td>11% (1.50E-01)</td>
</tr>
<tr>
<td>Campesino/Indigena</td>
<td>9% (6.78E-05)</td>
<td>5% (3.13E-01)</td>
<td>4% (3.93E-01)</td>
</tr>
<tr>
<td>Campesino/Mestizo</td>
<td>4% (4.19E-02)</td>
<td>24% (9.09E-09)</td>
<td>21% (2.73E-07)</td>
</tr>
<tr>
<td>Indigena/Mestizo</td>
<td>5% (4.7E-05)</td>
<td>20% (3.43E-05)</td>
<td>25% (1.47E-06)</td>
</tr>
</tbody>
</table>

The results of the $t$-test do, in some cases, support the assertion that the ethnic groups have somewhat different ancestries. In all comparisons, the mean proportion of West African ancestry among individuals who self-identify as Afro-Colombian is substantially and significantly greater than the mean for individuals who identify as one of the other ethnicities. However, this does not mean that every Afro-Colombian individual has more West African ancestry than every individual from another ethnic group. In fact, the distributions in Figure 3-7 illustrate that a number of individual self-identifying as Afro-Colombian have less West African ancestry than the average Mestizo individual. The other major trend in these results is that Mestizo individuals have, on average, more European ancestry than any of the other groups. Despite these differences in mean, all of the ethnic groups overlap at least one other ethnic group through half or more of their distribution. This directly contradicts any categorical differences among the ethnic groups.
Figure 3-7. Distribution of autosomal ancestry estimates by ethnicity. Point estimates of ancestry proportions for genomic West African, European, and Indigenous American ancestry based on MLE.
The distribution of autosomal ancestry in the sample as a whole differs markedly from that predicted from the NRY results. The BGA of the population based on the MLE results from the autosomal markers is 11.2% West African, 39.5% European, and 49.3% Indigenous American. Based on Y-haplotypes, the BGA is approximately 18.7% African, 57.7% European, and 23.6% Indigenous American. Table 3-2 shows these results for each ethnic group.

Table 3-2. Comparison of ancestry estimates for each ethnic group based on autosomal and NRY markers. Pairs of estimates in bold indicate an NRY estimate more than one standard deviation away from the mean estimate for the autosomal markers in that ethnicity.

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Autosomal</th>
<th>NRY</th>
<th>Autosomal</th>
<th>NRY</th>
<th>Autosomal</th>
<th>NRY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined</td>
<td>11.2%</td>
<td>18.7%</td>
<td>39.5%</td>
<td>57.7%</td>
<td>49.3%</td>
<td>23.6%</td>
</tr>
<tr>
<td>Afro-Colombian</td>
<td>48.1%</td>
<td>62.5%</td>
<td>17.1%</td>
<td>25.0%</td>
<td>34.8%</td>
<td>12.5%</td>
</tr>
<tr>
<td>Campesino</td>
<td>12.6%</td>
<td>25.0%</td>
<td>21.3%</td>
<td>41.7%</td>
<td>66.0%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Indigena</td>
<td>3.7%</td>
<td>7.7%</td>
<td>26.2%</td>
<td>38.5%</td>
<td>70.1%</td>
<td>53.8%</td>
</tr>
<tr>
<td>Mestizo</td>
<td>8.8%</td>
<td>15.6%</td>
<td>45.7%</td>
<td>65.6%</td>
<td>45.5%</td>
<td>18.9%</td>
</tr>
</tbody>
</table>

An interesting pattern emerges based on this comparison where both West African and European ancestry is consistently overestimated and Indigenous ancestry underestimated when using the Y haplotypes. Unfortunately, it is not possible to directly compare the Y and autosomal estimates using a statistical test because the Y estimate places each individual into a single ancestry category while the panel of autosomal markers apportions ancestry for each individual. In other words, Y haplotypes estimate every man to be 100% European, 100% African, or 100% Indigenous Americans. To estimate the degree to which the NRY and autosomal estimates differ within the ethnic groups and in the full population, the standard deviation of the autosomal estimates was computed for each ethnic group and cases where the difference between the NRY and autosomal estimates were greater than one standard deviation were considered significantly different.
Significant differences between the estimates were identified in the Mestizos and Campesinos for both European and Indigenous ancestry and in the Campesinos for West African ancestry. When the Campesinos and Indigenas are grouped together, the results are the same as for the Campesinos alone. The lack of significant findings in the Afro-Colombians may be due to the small sample size and the broad distribution of ancestry in this ethnic group which inflates the standard deviation. In the Campesino and Mestizo ethnic groups, the substantially larger proportion of European ancestry in the Y-chromosomes relative to the autosomes suggests male-biased European gene flow – a pattern that has been well established in the research summarized previously.

Ancestry, Ethnicity, and Phenotypes

Definitions of ethnic groups are inherently difficult things based on the comingling of ideas regarding ancestry, culture, language, politics, social status, physical appearance, and untold other variables. Although the above results indicate that there are some relationships between genomic ancestry and ethnicity, the complexity of ethnic designations raises the question of whether or not there may be a better predictor of an individual’s ethnic identity. Although the ethnic categories reported in this research are self-identified, it is not hard to imagine that an individual’s self-identity will be shaped in large part by how others perceive him or her. Phenotypes such as skin pigmentation may contribute to how individuals are described and describe themselves. Figure 3-8 shows the distribution of skin pigmentation (Melanin Index, M) by ethnicity where a higher M index indicates darker skin pigmentation.
Figure 3-8. Distribution of Melanin Index by ethnicity. Mean values for the three ethnic groups are 37.2 for Mestizos, 46.0 for Campesinos and Indigenas, and 54.9 for Afro-Colombians.

Comparing the distribution of the Afro-Colombians to the other two groups in terms of M index and West African ancestry (Figure 3-7) supports the assertion that skin pigmentation may be influencing self-description in this population. While there were a number of Afro-Colombian individuals with less West African ancestry than the average Mestizo individual, there are no Afro-Colombians who are lighter-skinned than the average Mestizo. Welch’s t-tests support a significant mean difference in skin color between Afro-Colombians and the combined Campesino/Indigena group (M difference = 8.9, \( p = 0.035 \)) and an even larger difference with Mestizos (M difference = 17.7, \( p < 0.001 \)). Interestingly, there is also a mean difference between the Mestizos and Campesinos/Indigenas (M difference = 8.8, \( p < 0.0001 \)).

These results are interesting but may be confounded by the relationship between skin pigmentation and ancestry. Because darker skin pigmentation is correlated with increasing non-European ancestry across this Colombian sample (\( R^2 = 0.385, p << 0.0001 \)), the association
between skin pigmentation and self-described ancestry could be due to a shared correlation with BGA. To determine if this is the case, skin pigmentation was regressed onto European, West African, and Indigenous American autosomal ancestry and the residual values were reevaluated using Welch’s $t$-tests and plotted in Figure 3-9. This minimizes the relationship between ancestry and skin pigmentation so the relationship between the residual melanin variation and self-described ancestry can be investigated.

![Figure 3-9](image)

Figure 3-9. Distribution of ancestry-adjusted residual melanin by ethnicity. After correcting for ancestry, the ethnic groups show more overlap in distribution of residual M with mean values of -0.03 for Mestizo, 2.47 for Campesino/Indigena, and 0.47 for Afro-Colombian individuals.

Adjusting for ancestry notably alters rank order of the means among the ethnic groups with the Campesinos and Indigenas now having the highest mean melanin value. This change is due to the unequal distribution of ancestry and skin pigmentation among the ethnic groups. After adjusting for the relationship between skin pigmentation and ancestry, results of the $t$-tests for mean differences in residual M index by population were no longer significant between the Afro-Colombians and the other two groups. This does not indicate that skin pigmentation does not
factor into how individuals self-report ethnicity, rather it suggests that skin pigmentation is not a better predictor of self-reported ethnicity than individual BGA. This assessment is also based on the limited sample size available in this study. In contrast to the results with the Afro-Colombians, the mean difference for residual M index between the Campesino/Indigena group and the Mestizos was still significant ($p = 0.008$). This might indicate that individuals with lighter skin are more likely to report as Mestizo, but the degree to which skin pigmentation is a factor in individual’s self conception is unknowable from this data.

**Conclusions**

Colloquial understandings of the terms Mestizo, Afro-Colombian, and Indigena all relate to presumed ancestors who are European and Indigenous American, African, and Indigenous American, respectively. While the results of the ancestry estimates could be viewed as supporting these classifications, the tremendous overlap in ancestry undercuts any such groupings based in biological typology. There are statistically significant differences in mean ancestry between Afro-Colombians and all other groups and between the Indigena/Campesino and Mestizo groups. However, there is also substantial overlap among all three of the groups on each ancestry distribution. While, on average, the Mestizos have the most European ancestry, the Afro-Colombians have the most West African ancestry, and the Campesinos and Indigena have the most Indigenous American ancestry, in each case, some members of the other ethnic groups overlap these means. As a result, it is impossible to accurately group individuals as they self-identify using either the NRY or the autosomal BGA estimates. The history of admixture in Colombia is deep and broad, and the boundaries between the populations are clearly permeable such that individuals with similar BGAs could belong to more than one group.
One aspect of this history can be clearly documented, however. Consistent with previous reports in admixed – and putatively unadmixed – populations throughout the Americas, the proportional ancestry estimated from NRY and autosomal markers are not equal. The distributions suggest European-male biased gene flow in Colombia, particularly in the Campesino and Mestizo groups. In many ways, the genetic similarities among socially distinct groups, as seen among the ethnic groups in this Colombian sample, is the overarching motif of stories of recent admixture in the Americas. The persistence of these classifications may have less to do with biogeographic ancestry and more to do with the political histories of the country and possibly associations between particular phenotypes such as skin pigmentation and putative ancestral populations.
References


Chapter 4

Testing Selection-Nominated Candidate Genes

In Chapter 2, a variety of methods were employed to identify signatures of selection in the genomes of Indigenous American individuals. Care must always be taken when attempting to characterize past selection events as there are so many unknowable factors that could influence the genomic distributions and undetectably confound the results. As these methods are all based on an empirical distribution, there is no statistical means of determining if these signatures of selection are indicative of true selection, drift, or simply a statistical artifact. The outlier based approach employed in this research, despite its many advantages, cannot perfectly differentiate between genes undergoing selection and those evolving neutrally, unless the cutoff value for the tail of the distribution is precisely equal to the proportion of SNPs or windows undergoing selection [1]. Of course, there is no a priori means of determining what proportion of the genome is or has been under selection or differentiating between regions of the genome subject to unusually high rates of drift. As a result, there are likely a number of genes that appear to be under selection when they are evolving neutrally.

Even in the case of the selection-nominated candidate genes that have actually undergone selection in a particular population, this does not guarantee that the favored polymorphisms contribute to population-level differences in pigmentation. Many of the genes involved in the melanin pathway are also involved in the development, transport, and structure of other cellular products. For example, the gene \textit{EGFR} is an epidermal growth factor receptor which has been implicated in a wide range of functions including proliferation, differentiation, motility, and survival in many types of cells, most of which do not produce pigment [2]. Variation in \textit{EGFR} could result in many phenotypic changes that may be advantageous or disadvantageous such that selection may be acting on this phenotype and not pigmentation.
Hundreds of papers have been published showing evidence for directional selection at a single gene of interest or reporting on genome-wide screens for selection. Very few papers, however, have continued on in an attempt to determine if any of these genes are contributing to within or among population differences in phenotype. This is understandable, particularly in the cases of genome-wide screens where the majority of available data is unlinked to phenotypes and in many cases the regions of putative selection do not contain genes with known functions or contain many genes with a variety of functions. In the cases of pigmentation genes, several studies have observed different signatures of selection in populations with differences in skin color. However, the genetic differentiation in pigmentation genes is weak evidence of functional differences, particularly considering the frequency of pleiotropy among genes related to melanogenesis. Until studies are done on individuals with quantified phenotypes, there is no justification to say that these allele frequency differences actually contribute to variation in pigmentation.

The two traditional methods for identifying genes contributing to complex phenotypes – linkage analysis and association studies – both have major limitations for identifying complex phenotypes like skin pigmentation that vary among populations. Linkage analysis has relatively low statistical power for effects that are modest in size and association analysis requires the genotyping of a large number of polymorphisms and shows decreasing power with increasing allelic heterogeneity. Admixture mapping (AM) is a method that exploits regions of linkage disequilibrium (LD) created by admixture to localize the genes responsible for traits that vary between the parental populations. In essence a form of association analysis, AM has the advantage of high statistical power for identifying causative genes where there are large allele frequency differences in the ancestral populations, as is expected to be the case for genes determining traits like skin pigmentation. AM also requires fewer SNPs than association studies to achieve the same sensitivity; however, it is intermediate between linkage and association.
analysis in terms of robustness to genetic heterogeneity and mapping resolution. The basic principles of AM are exploited in two ways in the following analysis.

The selection-nominated candidate genes will be tested using two related methods to determine what, if any role they play in contributing to the skin pigmentation difference between Indigenous Americans and Northern Europeans. There are two main challenges in identifying genes determining phenotypes that vary between populations. First, a large number of loci will show substantial allele frequency differences that have arisen since the populations diverged. The vast majority of these loci will not be related to the phenotype of interest. Second, there may be fixed differences between the populations at the causative loci and it is impossible to map genes without genetic variation. One way to overcome these hurdles is to examine the genes of interest in an admixed population where both the genes themselves and the backgrounds in which the genes occur will show the heterogeneity necessary to identify which variants are related to the phenotypic variation. The phenotypic variation in the admixed samples is also more likely to show the continuous variation expected for complex phenotypes such as skin pigmentation. From a purely logistical standpoint, the use of admixed individuals of Indigenous American and European ancestry is advantageous because admixed populations in the Americas today are more common than unadmixed populations of Indigenous Americans. The specific advantages of these approaches will be discussed in more detail in the following section.

Methods

Admixed populations were analyzed using two methods that exploit the unique advantages of these groups. The resulting associations between the selection nominated candidate genes and Indigenous American-specific changes in skin pigmentation revealed for the first time regions of the genome that responded to the high UVR of the New World tropics. This
research employs two complementary methods of assessing the relationship between variation in skin pigmentation and the selection-nominated candidate genes. ADMIXMAP allows for finer scale analysis by imputing the likely ancestry of specific regions of the genome. However, because the available data was genotyped using different AIMs panels, the full data set cannot be combined into a single set for analysis in ADMIXMAP. In contrast, non-parametric analysis of variance (ANOVA) lets the full data set be combined for greater statistical power.

Admixture Linkage Analysis

ADMIXMAP detects admixture linkage between skin pigmentation and these haplotypes in admixed individuals. Admixture linkage analysis exploits the gametic disequilibrium that results from recent admixture making it particular useful for investigating traits such as skin pigmentation which vary between populations. First described by Chakraborty and Weiss as a method for identifying genes linked with diseases [4], this method makes use of the fact that affected individuals will have a higher proportion of alleles derived from the high-risk parental population at the functional locus than unaffected individuals. Figure 4-1 is a simplified illustration of an admixed pedigree showing one pair of chromosomes for each individual.
In the first generation, unadmixed individuals from two populations which have been genetically isolated for an extended time are mating again for the first time. As a simplifying assumption, consider all loci on the chromosomes have an $F_{st}$ of 1 between the two populations meaning that alternate alleles are fixed in the populations. Individuals from population 1 (whose chromosomes are shown in blue) have lighter skin than individuals from population 2 (whose chromosomes are shown in red). Analysis in the first generation (I) would show that all of the alleles on the red chromosomes are associated with darker skin pigmentation than all of the alleles on the blue chromosome simply because those alleles occur more often in individuals from population 2 who have darker skin on average. In the second generation, individuals with one parent from each of the populations will be heterozygous at all loci and presumably have intermediate skin pigmentation. All of the alleles on the red chromosomes will still show association with darker skin pigmentation because they have all cosegregated with the causative allele. Admixture creates large regions of linkage disequilibrium (LD) as loci that were randomly
associated in the ancestral populations are now associated by virtue of occurring on the same chromosome passed into the admixed population. In generation III, the linkage disequilibrium will start to break down as recombination occurs during meiosis and individuals in generation II pass on chromosomes bearing a combination of alleles from both ancestral populations. In this generation, only very broad regions of the genome could be shown to be linked to the phenotype. For example, if individual III(a) has darker skin than individual III(b) then it could be established that the gene associated with darker skin pigmentation in population 2 must be in the top half of the chromosome because that is a region where individual III(a) has inherited a region of the chromosome from population 2 but individual III(b) has not. Over several more generations, linkage disequilibrium will break down even more and the regions that can be could be shown to be linked to the phenotype will be narrower and narrower. This linkage is established by comparing locus-specific genetic ancestry in individuals with the trait (labeled “cases” in the figure) to those who do not have the trait. Regions where cases but not controls show ancestry from the affected parental population are likely to be linked to the phenotype. In this instance, the dotted line marks a region of the chromosome where all cases have inherited at least one of the two chromosomal segments from ancestral population 2 and all of the controls have inherited both chromosomal segments from population 1. This is an indication that the causative gene is somewhere in this region.

For admixture linkage analysis to be successful, several generations must have elapsed for the large-scale LD caused by the initial admixture event to break down or else the regions of the genome identified as containing the causative gene will be so broad as to be completely uninformative. In contrast, if too many generations have passed, the LD will have broken down to the extent that many more markers will have to be genotyped to ensure coverage of each haplotype. When the admixture-induced LD breaks down completely, adjacent loci are equally likely to contain alleles from populations 1 or 2. The sum of intensities (τ) can be calculated for
the loci to determine if there is sufficient ancestry information in the AIMs genotyped to calculate locus-specific ancestry. The value of \( \tau \) is determined by the number of times the ancestry of a chromosome changes in a given length of the chromosome and is equal to the number of generations since the beginning of admixture in the population. [5] If the sum of intensities is too large, admixture linkage analysis becomes no more powerful than an association study.

Admixture mapping – also referred to as mapping by admixture linkage disequilibrium (MALD) – had been used in the identification of candidate genes for susceptibility to multiple sclerosis [6], hypertension [7], kidney disease [8], and other disorders. Additionally, this methodology has been used to study normal variation where a phenotype, such as skin pigmentation, varies with genomic ancestry in admixed populations [9-12]. AM has been likened to an experimental cross between inbred strains of mice that differ at a trait of interest [13]. However, this analogy is imperfect because admixed human populations are not the descendents of inbred strains that share no alleles (i.e. have an \( F_{st} \) of 1), nor is the investigator able to dictate which pairs of individuals mate over many generations as one could with lab animals. However, these problems can be overcome statistically by estimating locus-specific ancestry using Ancestry Informative Markers (AIMs) in linkage disequilibrium with the loci of interest. This is similar to the method developed by Pritchard and Donnelly [14] to control for population stratification in standard association studies. A large panel of AIMs genotyped across the genome also allows estimation of proportional ancestry for regions of the genome based on a multipoint analysis comparing the observed variation to the allele frequencies of the parental populations. While any single locus may not be sufficiently informative, a denser panel of AIMs will enable a correction for differing individual genetic histories and admixture levels. [15]

These statistical adjustments can be calculated relatively easily using the program ADMIXMAP [5, 16], which was designed specifically for use in this type of research. The major advantage of which is that it allows for the investigation of loci that vary between populations to
the exclusion of variation within the parental populations and requires orders of magnitude fewer markers than a standard association study [17]. Because recent admixture creates extended regions of LD, only the AIMs previously genotyped for ancestry estimation and a set of markers with large $\delta$ in each selection-nominated candidate gene were needed for this analysis. Longer times since admixture would require the use of a denser marker panel.

ADMIXMAP addresses a number of challenges that are inherent in the use of admixed samples in linkage analysis. The first is related to the difference between laboratory controlled admixture between strains of inbred model organisms and naturally occurring human populations. In the lab, the ideal proportion of ancestry (50%) is maintained by mating two individuals with 50% ancestry together. Over several generations, there will be an increasingly broad distribution of ancestry in the population due to independent assortment, but the distribution will always be centered on 50%. In human populations, admixed individuals may selectively mate with other admixed individuals of a particular ancestry proportion or individuals from the parental populations. These individual differences in admixture history lead to stratification within the sample which can confound associations between alleles more common in a particular subpopulation and phenotypes more common in that subpopulation. This can be addressed either by fitting a generalized linear model of individual admixture proportions to the phenotype (as will be discussed in the following section on analysis of variance) [16] or by comparison of locus-specific ancestry to expected locus ancestry based on parental BGA. The latter method is employed in ADMIXMAP which also allows for a multipoint analysis of locus ancestry using multiple adjacent markers which compensates for imperfectly informative (not fixed) polymorphisms between the parental
populations. The statistical power of this method is increased with more informative (higher $F_{st}$) and more densely spaced markers. As a result, multiple markers spaced in and around the locus of interest increases the accuracy of the local ancestry estimate and the power to detect association with the trait. [5]

A final consideration is that unlike the case of inbred strains, the precise ancestral allele frequencies for an admixed population cannot be known. Instead, ADMIXMAP calculates likely ancestral frequencies based on the combination of the admixed sample and allele frequencies in unadmixed, contemporaneous populations. The possible changes in allele frequencies in the intervening generations between members of the ancestral populations forming the admixed populations and the generation of unadmixed descendents that was sampled to obtain allele frequency as well as the possible sampling error that is incurred when sampling the contemporary proxy parental populations are factored into the ancestral allele frequencies through the use of a Bayesian model. [5]

The ancestry for each allele (maternal and paternal) at each locus is modeled on a Dirichlet distribution of ancestry in the sample, $\alpha$. The stochastic distribution of ancestry across the chromosome due to recombination is generated and modeled with a Markov process which at any locus maintains the Dirichlet distribution of ancestry. This processed is based on a number of Poisson distributions of the likelihood of a change in ancestry state equal to the number of ancestral populations. The Markov process is “memory less” meaning that the ancestry at each locus is estimated independently of the adjacent loci as would be expected under independent assortment. In this way, an individual’s ancestry for each allele can be modeled but the state of that allele will be
determined by the ancestral allele frequencies. [16] Figure 4-2 demonstrates the interrelationship between these parameters.

Figure 4-2. Interrelationships between parameters estimated by ADMIXMAP. Each individual’s genotype is determined by the paternal and maternal locus ancestries which are in turn determined by paternal and maternal admixture. Paternal and maternal admixtures are determined by the population-level distribution of admixture assuming random mating. For each locus, the specific allele carried by the paternal or maternal locus of a given ancestry will be determined by the frequency of each allele in the ancestral populations. The trait measurement for a given individual is also influenced by parental admixture, the regression parameters which describe the relationship between ancestry and the phenotype in the absence of locus-specific effects, and covariates. Figure from [16].

Having estimated the ancestry states of the alleles at each locus, a model can then be fit for the dependency of the phenotype of interest on the ancestry states across the genome. In the
case of a continuous variable like skin pigmentation, this is a linear regression. The regression is fit based on a Markov chain-Monte Carlo (MCMC) simulation of the posterior distribution of ancestry for all unobserved loci conditioned upon the observed genotypes. The phenotype of interest is then tested for association with each locus based on a likelihood based test calculated as the alternative to a null hypothesis which is a generalized linear model including the observed and expected genotypes conditioned on the distribution of ancestry in the genome, informativeness of the marker for ancestry, mean admixture proportions of the parental gametes, any available covariates, and the log odds ratio of the phenotype. [18]

Essentially, this process results in the detection of regions of the genome with an excess of Indigenous American ancestry relative to the population mean in individuals with darker skin pigmentation. Such an excess is illustrated in Figure 4-3. Because the goal of the analysis is not simply to assess directly the relationship between the markers tested and the trait, but rather to determine the relationship between ancestry in a region and phenotype, the precise markers are not as important as their ability distinguish between ancestry states. Regions where individuals with darker skin pigmentation have one or two copies of the alleles descended from Indigenous American ancestors and no copies of the ancestral European alleles are likely to contribute to population-level differences in skin pigmentation. If individuals 1 and 2 have darker skin than individual 3, the region of the genome highlighted by the box would show significant results because the individual with darker skin pigmentation show more Indigenous American Ancestry in this region than the lighter-skinned individual.
Figure 4-3. Illustration of individual ancestry estimates along a chromosome. The dotted box indicates a region of high Indigenous American ancestry in individuals 1 and 2 and low Indigenous American ancestry in individual 3. Figure adapted from [19].

**Non-Parametric Analysis of Variance**

The predominant limitation of admixture mapping for the current analysis is that it is not robust to large quantities of missing data and this sample contains subsamples genotyped for three separate sets of markers. To determine the relationship between skin color and the genotypes at the genes of interest, a non-parametric analysis of variance, the Kruskal-Wallis test, was performed on the samples. The Kruskal-Wallis test must be employed because a number of the genotypes are not normally distributed. Normality was tested using the Shapiro-Wilks’ W test for neutrality in the stats package of R [20]. The Kruskal-Wallis test quantifies the relationship between skin pigmentation and the genotypes of interest but because of the stratification that is
created in the population by admixture, it is essential to adjusted for the relationship between ancestry and melanin. [21]

Rather than estimating locus-specific ancestry which requires complicated modeling as in ADMIXMAP, genome-wide biogeographic ancestry (BGA) is included as a proxy for locus-specific ancestry in the analysis of variance. In an individual, the likelihood of the alleles at any given locus having been inherited from one parental population is equivalent to the total proportion of the genome descended from ancestors in that population, the autosomal BGA as described in the previous chapter. BGA can be calculated in a number of ways including maximum likelihood estimation (MLE) [22], a Bayesian approach [23] as implemented in the program STRUCTURE [24], or a combination of these approaches as applied in the program ADMIXMAP [5, 16]. These three approaches have been found to be equally robust [25] so the MLE is employed for the analysis of variance.

MLE is calculated by calculating for each individual the most likely value of \( m \), the proportion of ancestry from population 1, such that the following three equations are maximized.

\[
\begin{align*}
\Pr(A_1A_1) &= (m_{p1} + (1 - m)p_2)^2 \\
\Pr(A_1A_2) &= 2 \times [m_{p1} + (1 - m)p_2] \times [m_{q1} + (1 - m)q_2] \\
\Pr(A_2A_2) &= (m_{q1} + (1 - m)q_2)^2
\end{align*}
\]

where \( p_1, p_2, q_1, \) and \( q_2 \) are the allele frequencies in the unadmixed ancestral populations. These equations, derived from Hardy-Weinberg equilibrium, show that the probability of an individual having each genotype is a function of the proportions of ancestry from each ancestral population which are represented as \( m \) and \( 1 - m \). [26] MLE was calculated for each individual using a Perl script written by Vibhor Sonpar.

At the population level, an average Indigenous American ancestry of 52% means that, in the absence of selection since the time of admixture, 52% of the alleles at the loci
of interest are on average descended from the Indigenous American parental population and 48% from the European parental population. Because darker skin pigmentation is associated with increased Indigenous American ancestry, the AIMs are essential to control for genomic ancestry so that the effect of the selection-nominated candidate genes can be accurately assessed. To identify cases where there is an excess of Indigenous American ancestry and that excess is associated with differences in skin pigmentation, the overall relationship between skin pigmentation and Indigenous American ancestry must be controlled for. Effectively, this adjustment eliminates the proportion of skin pigmentation variation that is due to the genetic “background” so that the proportion of the variation specifically related to the genes under consideration can be properly evaluated. To do this, a regression is taken of skin color on ancestry. There is a strong correlation between increased Indigenous American ancestry and darker skin pigmentation (larger Melanin Index) in the combined sample (Pearson’s Adjusted $R^2 = 0.37, p < 0.00001$) as can be seen in Figure 4-4. The residual M values from this regression are then used for future analysis. As seen in the second graph of Figure 4-4, the Ancestry Adjusted Melanin Index values are no long correlated with Indigenous American ancestry (Pearson’s Adjusted $R^2 = -0.002, p =1$).
Figure 4-4. Relationship between skin pigmentation and Indigenous American ancestry in sample. The top plot clearly indicates the strong correlation between Melanin Index and individual estimates of Indigenous American ancestry (Pearson’s Adjusted $R^2 = 0.37$, $p < 0.00001$). The bottom plot shows the Ancestry-Adjusted Melanin Index values taken from the residuals of the above regression regressed against Indigenous American ancestry (Pearson’s Adjusted $R^2 = -0.002$, $p = 1$).
Genotyping SNPs in Selection Nominated Candidate Genes

To afford the maximum ability to detect patterns of association between the Indigenous American ancestry in the genes of interest and darker skin pigmentation, SNPs were selected in each gene that show high allele frequency differences (δ) between the European and Indigenous American populations in the Mega sample. SNPs with larger δ are more informative for the ancestry of a region of the genome so deviations from the expected proportion of Indigenous American ancestry in dark-skinned individuals can be more readily detected.

For each of the 14 selection-nominated candidate genes identified in Chapter 2, one to three SNPs were genotyped for a total of 29 SNPs in the selection nominated candidate genes plus SNPs in SLC24A5 and MATP. Genotyping was combined into a single multiplex reaction which was run at the University of Minnesota BioMedical Genomics Center using the Sequenom iPLEX Gold assay on the MassARRAY platform. The multiplexing places some limits on the SNPs that can be combined due to proximity to one another on the genome, local tertiary structure of the DNA, and an assortment of other causes. All samples were prepared for processing at the Anthropological Genomics Lab at Pennsylvania State University with the exception of the Albuquerque samples which were prepared for genotyping and, in compliance with University of New Mexico Human Research Review Committee regulations, preliminarily amplified in the Department of Anthropology at UNM. This genotyping and all analyses were performed under Pennsylvania State University IRB approval (IRB #28267). Two SNPs failed in the genotyping while the remainder had call rates over 97%.

The previously identified SNPs in SLC24A5 (rs1426654) and MATP (rs16891982) were included in the analysis because they have previously been shown to be associated with skin-pigmentation differences between European and Africans [27]. The distribution of alleles in these is such that they appear to have mutated and spread nearly to fixation in Europe making
them reliable markers for, and probably cause of skin pigmentation difference between, any European and non-European populations. As a result, they must be included in this analysis because they are likely to influence skin pigmentation in this admixed sample. As with adjusting for the influence of genome-wide ancestry on skin pigmentation values, including the genotypes of the SLC24A5 and MATP SNPs in the analysis will remove as much variation as possible so that the variation related to the selection-nominated candidate genes and their function in the darker pigmentation of the Indigenous American populations.

While it is not anticipated that the tested SNPs in the selection-nominated candidate genes will be the functional SNP, they will indicate if the gene of interest has a higher proportion of Indigenous American ancestry in individuals with darker skin pigmentation, in other words, shows admixture linkage. Although it is not the specific goal of the study, this research has the potential for identifying SNPs that account for lighter skin in Europeans or that may be functional in East Asians as well as Indigenous Americans due to shared haplotypes between these populations. This is because the SNPs, by virtue of capturing the variation related to Indigenous American ancestry in the genes, will also capture the inverse, the variation due to European ancestry.

Samples

The 515 individuals included in this portion of the study have a range of levels of admixture between Indigenous American and European parental populations. They reside predominately in regions of the Americas with high ground-level exposure to UVR and where indigenous individuals have darker average skin pigmentation than populations in Northern Europe and East Asia. The geographic positions of each of the sampling locations can be seen in
Figure 4-5. Although current residence in high-UVR climates does not guarantee that their Indigenous American ancestors were exposed to the same conditions, it is a reasonable proxy.

The sample is compiled from four populations, each of which were sampled under somewhat different protocols as will be discussed. Due to these differences as well as the different latitudes at which the collections were performed, the samples will be analyzed both together and individually. In all cases, the samples have been previously genotyped for panels of AIMs to estimate autosomal genomic ancestry proportions and individuals with more than 10% West African and/or East Asian genomic ancestry were removed from the sample to minimize the influence of alleles other than those commonly found in Indigenous American and European populations. Additionally, skin color was measured for all participants using reflectometry on the proximal medial portion of the arm with multiple measurements across both arms averaged to
generate a single value. This location was chosen to best capture constitutive skin pigmentation – the basal amount of melanin produced by the body due to genes – not the effect of sun exposure on the skin. The amount of time spent in the sun and the amount of melanin produced in response to sun exposure vary substantially between individuals. This variability will confound attempts to identify genes related to constitutive pigmentation which are not necessarily overlapping with genes related to melanogenic response to UV exposure. The sample sizes discussed are for the number of individuals included in the final analysis, not the number of individuals sampled or the number for whom genotyping was attempted.

Guerrero, Mexico Samples

Ninety-five individuals included in this analysis were collected in the city of Tlapa in the state of Guerrero in southwestern Mexico [29]. Subjects were collected at the market in Tlapa under appropriate informed consent (PSU IRB #00M0525-00) by Gerrardo Gutiérrez and collaborators. In addition to a peripheral blood draw for DNA analysis, volunteers were asked to respond to a series of questions regarding their ethnicity and the ethnicity of their parents with the majority of individuals identifying as members of one or more indigenous groups (predominantly Mixtec, Nahua, and Tlapanec) or Mestizo. Tlapa is located in a mountainous region and is home to a large concentration of the indigenous population of the state of Guerrero. The market where the subjects were ascertained is the primary market in the region with a great number of people traveling from the surrounding areas.

Constitutive skin pigmentation was measured by reflectance spectroscopy with a DermaSpectrometer (Cortex Technology, Hasund, Denmark) following a standard protocol [30] for all individuals. The DermaSpectrometer is a narrow-band spectrophotometer with a green diode centered on 568 nm and a red diode centered on 655 nm. The melanin index (M), is
computed from the reflectance of the red diode and is well documented as an accurate measure of constitutive pigmentation levels [31]. The melanin index is computed as the base-ten log of 1/% red reflectance. Because darker skin absorbs more of the red light, individuals with more melanin (darker skin) will have a higher M. The range of M in this sample is 23.6-57.7, with a mean of 46.4 ± 4.9. For comparison, a sample of European Americans from State College, PA, has a mean M of 30.5 ± 3.0 and an African-Caribbean population averages 57.0 ± 5.2 [9]. The distribution of M values among the Guerrero and in the full sample can be seen in Figure 4-6.

Individual BGA was calculated for the Guerrero individuals using MLE on a panel of 24 autosomal AIMs selected to distinguish between Indigenous American, Spanish, and West African parental groups [11]. This panel was generated using references populations from Valencia, Spain; Maya, Cheyenne, Pima, and Pueblo Indigenous Americans; and individuals from Nigeria, Central African Republic, and Sierra Leone. This panel includes both SNP and insertion/deletion polymorphisms. All markers have a minimum allele frequency difference (δ) of 0.3 between two parental populations. Summed δ values are 11.7 for European-Indigenous American, 14.4 for European-West African, and 16.6 for Indigenous American-West African discrimination. Of the 24 markers, 19 markers were informative for distinguishing between European and Indigenous American ancestry contributions which is the primary axis of interest for these populations. The remaining 5 markers were used to identify individuals with West African ancestry. These individuals were removed from the sample. Allele frequencies and details about this panel of AIMs can be found in Appendix C. This sample contains a number of individuals with a large proportion of Indigenous American ancestry with a range from 83-100% with a particularly high average of 98%. As was noted in the original paper and can be seen from the discussion of BGA among indigenous populations in South America, this is an unusually low level of admixture for most populations in the Americas. The distribution of Indigenous American ancestry for the Guerrero and all other samples can be seen in Figure 4-6.
Figure 4-6. Distribution of Ancestry and Melanin. For optimal results from the admixture linkage analysis performed in this study, the distribution of both BGA and the outcome variable (M) should be normally distributed with the ancestry distribution having a mean near 50%. The combined sample is ideal for this type of analysis.
San Luis Valley, Colorado Samples

Two U.S. samples were included in the analysis. The first is 180 Hispanics from the San Luis Valley, Colorado which were collected as part of the San Luis Valley Diabetes Study [12]. The San Luis Valley is a geographical region stretching from southern Colorado into northern New Mexico. Hispanic individuals in this region frequently self-identify as Spanish Americans rather than Hispanic because they trace their ancestry to individuals living in the region prior to annexation by the U.S. in 1848. Interestingly, the above referenced study found that individuals who self-identified as Spanish American had less Indigenous American ancestry than those who identified as Mexican-American. However, self-identification was not considered as a variable in the present study. Individuals were transferred to Pennsylvania State University under IRB approval (ORC# 00M0453). Neither skin pigmentation nor ancestry estimates varied between the diabetic and non-diabetic individuals collected as part of this study so both were included in the analysis. A number of diabetes and metabolic syndrome phenotype were collected as well as skin pigmentation using a Photovolt model 575 spectrophotometer (Photovolt Instruments Inc., Minneapolis, MN) which takes measurements using glass filters at three intervals in the visible spectrum – amber (650 nm), green (550 nm), and blue (450 nm). From these values, Lightness (L*) can be calculated from the green measurement (g) using the formula

\[ L^* = 116 * \left( \frac{g}{100} \right)^{1/3} - 16. \]

To allow these samples to be compared with the other samples under analysis, the L* values had to be converted to M values. This was done following Shriver and Parra [30] using the equation

\[ M = 300.25 - 63.646 \ln (L^*) . \]

The relationship between these values is imperfect with an \( R^2 = 0.9617 \) for medial arm measurements. To minimize the influence of this conversion, the L* values were used in any
analyses where the San Luis Valley sample was not combined with other samples. After conversion, the distribution of M values in this sample ranges from 30.2-51.4 with a mean of 37.7 ± 3.4.

Ancestry was calculated for this sample using the same autosomal AIMs panel as the Guerrero sample. The distribution of Indigenous American Ancestry for the sample was 0-96% with a mean of 34.0%. Compared to other Hispanic populations in the U.S., the broad distribution of values is expected as well as the relatively low proportion of Indigenous American ancestries. A sampling of other Hispanic populations that have been assessed for BGA shows point estimates of 29% Indigenous American ancestry in a group of Mexican Americans in Arizona, 34.1% in Puerto Ricans from New York City, 40% in Mexican Americans in California [32].

Albuquerque, New Mexico Samples

The other U.S. population included in this sample is 67 self-identified Hispanic and Indigenous American students collected at the University of New Mexico by Yann Klimentidis under supervision by the University of New Mexico Human Research Review Committee (HRRC# 04-087, re-accessed by Ellen Quillen under HRC# 09-319) [33]. A number of individuals in this sample also self-describe as Spanish American and the use of the term here has its roots in the same historical events. However, recently the use of the term has increased among individuals who wish to distinguish themselves from recent Mexican immigrants. The purpose of the study for which the samples were originally collected was to identify interrelationships between BGA, self-identified sub-ethnic designations (like Spanish American), pigmentation, and how individuals perceive the ancestry of others. As part of sample collection, DNA was collected via cheek swabs and skin pigmentation was collected using the DermaSpectrophotometer. Both
constitutive skin pigmentation from the medial arm and facultative skin pigmentation measures from the forehead were taken and calculated via the M index. Only the constitutive readings were included in this analysis. The distribution of M in this sample ranges from 23.8-45.2 with a mean of 35.2 ± 3.9.

Individual BGA was ascertained for these individuals using MLE based on a 176 SNP AIMs panel at DNAPrint, Inc. (formerly operating in Sarasota, FL). This panel contains markers for distinguishing between West African, European, East Asian, and Indigenous American populations [34]. These AIMs were selected for high values of δ, Fst, and LSBL between West Africans from Nigeria, Sierra Leone and Central African Republic, Europeans (self-reported “Caucasians”) from the U.S., East Asians from the Coriell cell repository and first or second generation Asian Americans from the U.S., and Indigenous Americans including Mixtec and Nahua from Guerrero. All AIMs have a δ of at least 0.4 between two of the ancestral populations with summed δ values of 56.6 between Europeans and West Africans, 59.7 between Indigenous Americans and West Africans, and 47.1 between Europeans and Indigenous Americans. Forty-one of these AIMs contribute substantially to distinguishing between European and Indigenous American Ancestry. Based on these markers, the BGAs for the New Mexico sample have a range of Indigenous American ancestries from 8-92% with a mean of 37.8%.

**Combined Samples**

The samples described above were combined with 173 unrelated individuals collected in Popayán, Colombia as described in the previous chapter. The individuals in this sample have a range of Indigenous American ancestry from 0-100% with a mean of 51.9%. When combined, the distributions of both skin pigmentation and Indigenous American ancestry in the admixed samples are broad and will be sufficient to detect admixture linkage between the haplotypes and
variation in skin pigmentation in these populations. The combined sample has a highly desirable
distribution of Indigenous American ancestry with a range from 0-100%, a nearly normal
distribution, and a mean of 52.4% which is close to the ideal average of 50% for admixture-based
analyses. In the total sample, the melanin index ranges from 22.7-65.3 with a mean of 39.3 ± 8.3.

Results

Genotype Distributions

Figure 4-7 shows the distribution of genotypes in the sample as a whole. Because the
allele frequency differences (δ) in the parental populations vary among the SNPs, the
distributions of the alleles will vary as well. In SNPs were there is a large δ, as with SLC24A5
where alternate alleles are fixed in Western Europeans and Indigenous Americans, the
distribution of alleles should mirror the ancestry frequencies. Since the population is
approximately 50% European, and 50% Indigenous American in origin, the genotype frequencies
should be 25%-50%-25% for SLC24A5 if assuming the population is in Hardy-Weinberg
equilibrium. However, this is clearly not the case for SLC24A5 which is one of seven SNPs
(rs1426654, rs11895982, rs16891982, rs7596929, rs11238349, rs1015361, rs12439639) that are
not in Hardy-Weinberg Equilibrium.
Figure 4-7. Distribution of Genotypes for Selection Nominated Candidate Genes. Many genotype distributions in the sample are non-normal. This non-normality dictates the types of analysis that must be used to analyze these results.
Where there is an excess of homozygotes, Hardy-Weinberg disequilibrium may indicate that there is substructure in the population. It is likely that there is substructure in the sample as a whole considering the different distributions and relationships of melanin index (M) and BGA in the four populations, as seen in Figure 4-8. Spearman rank coefficient adjusted $R^2$ values are 0.31 ($p < 0.0001$) for Colombia, -0.0106 ($p = 0.91$) for Guerrero, 0.09 ($p < 0.0001$) for San Luis Valley, and 0.13 ($p < 0.002$) for Albuquerque samples. In all cases except for the Guerrero, there is a significant correlation between skin pigmentation and ancestry. The low correlation in the Guerrero is likely due to the small amount of variation in both skin pigmentation and ancestry. With the broadest distribution of both M and ancestry and the strongest correlation between the two, Colombia will likely give the best results of any of the populations in individual analyses, as will be done with ADMIXMAP analysis, but the combined sample which has an even broader distribution and larger $R^2$ value in addition to the larger sample size should have the greatest power to detect relative small effects.
Figure 4-8. Regressions of melanin index (M) versus Indigenous American ancestry in four populations. Spearman rank coefficient adjusted R^2 values are 0.31 (p < 0.0001) for Colombia, -0.0106 (p = 0.91) for Guerrero, 0.09 (p < 0.0001) for San Luis Valley, and 0.13 (p < 0.002) for Albuquerque samples.

**ADMIXMAP Results**

Because ADMIXMAP does not respond robustly to large quantities of missing data, the four populations cannot be combined into a single analysis using this platform. The inclusion of the raw genotype data from the AIMs panels is essential for maximizing the accuracy of the locus-specific ancestry. As a result, only the two largest datasets – from Colombia and the San Luis Valley – were assessed independently using ADMIXMAP. For each data set, parental allele frequencies were included as provided in the original papers for the AIMs panels or, in the case of the SNPs tested in this work, as determined from the Mega sample. The allele frequencies for MATP and SLC24A5 were taken from dbSNP (ncbi.nlm.nih.gov/snp). These allele frequencies form the basis of the prior distributions used to estimate ancestry for each locus. A number of
results are supplied to determine the accuracy of the results and in both cases the number of iterations had to be increased until the scores were stabilizing by the ends of the run. A number of tests were performed including individual and population-level admixture and the results showed a high degree of concordance with the MLE point estimates for ancestry at both the population and individual level. Additionally, the program calculates the $\tau$ value which indicates the number of generations since the parental populations were unadmixed. For both populations, $\tau = 6$ generations, or approximately 150 years using a 25-year generations time. Although this seems like fewer generations than would be expected considering the population histories, it is similar to the 7 generations estimated for admixed individuals from Mexico city [35].

The most relevant of these tests for this work is the allelic association test where a linear regression was modeled between M and locus-specific ancestry as calculated based on the posterior distributions. This is a score-based test where values greater than 3 are considered significant. The $z$-scores (likelihood odds ratios, LOD) for the allelic association test can be seen in Figure 4-9. In addition to the allelic association test, an ancestry-association test was performed where the M index was regressed against the number of alleles from each ancestral population at each loci. Because the SNP density outside of the candidate gene regions was low, this test did not provide any additional information, but generally agreed with the results of the allelic association test.
Figure 4-9. Distribution allelic association Z-scores across genome in Colombian and San Luis Valley populations. The absolute values of the z-score (vertical axis) were plotted for the San Luis Valley Colombian results. The z-scores can be either negative or positive reflecting how the alleles were coded, so a value over 3 either positive or negative is significant. Z-values are plotted for the candidate SNPs which were typed in both the Colombian and the San Luis valley populations as well as the AIMs which are unique to each population. Genes are labeled that have a z-score exceeding 3. The horizontal axis is to scale and labeled with the cM location of the SNPs along the chromosome.
The ADMIXMAP results show the highest z-scores on chromosome 15 in a region that includes the genes PLDN, SLC24A5, and MYO5A. SLC24A5 is known to contribute significantly to skin pigmentation differences between European and West African populations with alternate alleles fixed in European and most non-European populations including Indigenous Americans [27, 36, 37]. The effect of this SNP (rs1426654) on skin pigmentation may be so strong that the SNPs for PLDN (rs12439639), MYO5A (rs12396, rs8026828, rs4776017), and RAB27A (rs11632529) may appear to be associated with variation in skin pigmentation due to linkage disequilibrium with SLC24A5. Although these SNPs are 2.58, 4.0, and 7.11 Mb away from SLC24A5, the long stretches of linkage disequilibrium created by the admixture process makes this distance well within a reasonable range to expect to see LD. If the adjacent SNPs were being swept along due to a strong effect in SLC24A5, the z-scores should decrease steadily with increasing distance from that gene. The z-scores for the region of chromosome 15 surrounding SLC24A5 are plotted in Figure 4-10.

Figure 4-10. Z-scores for allelic association on chromosome 15. The dotted line indicates the z-score threshold of 3. The SNPs are represented in order in this figure but the distances between them along the chromosome are not to scale.
Although the z-score for \textit{SLC24A5} is clearly the largest in the region for both populations, it appears that \textit{MYO5A} and particularly \textit{RAB27A} have independent peaks which may be an indication that these genes play a role in skin color variation in this sample. A larger sample size may be necessary to identify enough individuals having had recombination events between the genes which would allow the independent assessments of these loci.

The next largest z-scores in the analysis were found near \textit{MATP} which is another gene previously strongly associated with substantial European/non-European differences in skin pigmentation. That this gene would play a large role in skin pigmentation variation in this admixed sample is unsurprising. The next three largest scores were for \textit{ASIP}, \textit{ATRN}, and \textit{AP3D1}. Interestingly, these were only detected in the San Luis Valley sample and not the Colombian sample. There are a number of possible explanations for this. One is that with the slightly larger number of AIMs on these chromosomes, the locus-specific ancestry may be better estimated and corrected for in this population. If this is the case, these are a statistical artifact, not a true association. However, it could also be the case that there is regional variation in the effect. Finally, the gene \textit{PAX} has a z-score over 3 for both the Colombians and San Luis Valley samples.

\textbf{Power Calculations}

That \textit{SLC24A5} and \textit{MATP} were detected as significantly correlated with variation in M in both the Colombian and San Luis Valley ADMIXMAP analyses suggests that both have the ability to detect at least very strong effects resulting from skin lightening genes common in Europeans. However, to determine what effect sizes can be detected with these samples, it is necessary to calculate the statistical power of this test. Although there are several methods of determining the power for association studies, it is important to consider the unique statistical advantages of an admixture-based analysis and use a methodology specifically developed for
these research designs. The calculation of the statistical power was determined following McKeigue et al. [23] and Hoggart et al. [5]. The two main parameters that influence power for admixture linkage studies are sample size and the ancestry proportions of the sample (which are assumed to be normally distributed). The sample size needed for a given effect size ($\lambda$) can be calculated as

$$n = \left(\frac{Z_{1-a} + Z_{1-b}}{\lambda}\right)^2 \cdot \frac{1}{V}$$

where $Z_{1-a}$ and $Z_{1-b}$ are fixed at 4.27 and 1.28 such that the squared sum of these values is 30.8 which is the required information $n^*V$ to detect an effect of size $\lambda = 1$ with 90% power with a one-sided $p$ value of 0.00005. $V$ is the expected ancestry information from two gametes (chromosomes) with admixture proportions $\theta$. For a cross-sectional study of a continuously varying trait like pigmentation, $V$ can be calculated

$$V = \frac{1}{2} \cdot \theta \cdot \frac{(1 - \theta)}{\sigma^2}$$

where $\sigma^2$ is the residual variance. With the available sample size of 173 Colombians and an ancestry distribution of 52% Indigenous American/48% European, this population will have 90% power to detect genes with an effect size of at least 1.19 which corresponds to a risk ratio of 3.3. With a sample size of 180, but a less favorable ancestry distribution of 34% Indigenous American/66% European, the San Luis Valley sample has 90% power to detect polymorphisms with the slightly larger effect size of at least 1.22 which equals a risk ratio of 3.4.

SNP density will also play a role in the efficacy, but not strictly speaking the power of the analysis. This power calculation assumes that at least one marker is in LD with the causative allele. For a recently admixed population and markers in the genes of interest, this is a reasonable assumption. However, if instead of testing specific genes, one were to search genome-wide, more markers would be required.
Kruskal-Wallis Tests

As a comparison to the results from ADMIXMAP and to increase the sample size through the use of the full data set, analysis of variance methods were employed. When melanin values are regressed against the genotypes with no correction, nearly all of the SNPs tested show significant $p$-values (Table 4-1). This is to be expected since the SNPs were selected to be ancestry-informative so even a non-pigmentation SNP with a large $\delta$ would appear to be correlated with variation in skin pigmentation. Because a number of the genotype distributions were non-normal according to Shapiro-Wilks’ W test, the non-parametric Kruskal-Wallis test was computed in R. Sex was considered as a covariate but was found not to show a strong correlation with skin color in this sample and so it was not included in further analyses. In all cases, the MLE point estimates of individual Indigenous American ancestry was used as a covariate in the analysis. Adding West African ancestry as an additional covariate was considered, but it was not found to be significantly associated with skin pigmentation in this sample, likely because individuals with more than 10% West African ancestry were excluded from the sample. The $p$-values of all Kruskal-Wallis tests can be seen in Table 4-1. Following the inclusion of ancestry as a covariate, only $SLC24A5$ and $MATP$ show $p$-values significant after the Bonferroni correction for multiple testing.
Table 4-1. *P*-values for Kruskal-Wallis Tests. The values below represent the *p*-values for non-parametric analysis of variance (Kruskal-Wallis test, KW). Column label “No Correction” is the results for a KW test of M onto genotype. “W/ ancestry” is a KW test of M corrected for ancestry onto genotype. “+ w/ SLC24A5” is the same but using M corrected for ancestry and SLC24A5 genotype. “+w/ MATP” is as above with M corrected for MATP genotype as well. Haplotypes are M adjusted for all three parameters onto the phased haplotype for each gene containing more than one SNP. *P*-values less than 0.05 are italicized and those less than 0.0018 (Bonferroni correction value) are in bold. Asterisks (*) indicate that the SNP is not in Hardy Weinberg equilibrium.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene</th>
<th>Chr</th>
<th>Loc (Mb)</th>
<th>No Correction</th>
<th>w/ ancestry</th>
<th>+w/ SLC24A5</th>
<th>+ w/ MATP</th>
<th>Haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs11895982*</td>
<td>ADAM1 7</td>
<td>2</td>
<td>9.52</td>
<td>5.11E-16</td>
<td>0.0533</td>
<td>0.7552</td>
<td>0.4396</td>
<td>0.5995</td>
</tr>
<tr>
<td>rs7596929*</td>
<td>ADAM1 7</td>
<td>2</td>
<td>9.67</td>
<td>1.66E-12</td>
<td>0.1425</td>
<td>0.9344</td>
<td>0.6099</td>
<td></td>
</tr>
<tr>
<td>rs551573</td>
<td>POMC</td>
<td>2</td>
<td>25.22</td>
<td>2.36E-07</td>
<td>0.2095</td>
<td>0.5799</td>
<td>0.4191</td>
<td></td>
</tr>
<tr>
<td>rs12233134</td>
<td>POMC</td>
<td>2</td>
<td>25.24</td>
<td>1.88E-07</td>
<td>0.8371</td>
<td>0.4413</td>
<td>0.0996</td>
<td>0.1133</td>
</tr>
<tr>
<td>rs10186053</td>
<td>POMC</td>
<td>2</td>
<td>25.32</td>
<td>5.88E-05</td>
<td>0.6908</td>
<td>0.6191</td>
<td>0.3192</td>
<td></td>
</tr>
<tr>
<td>rs6748403</td>
<td>PAX3</td>
<td>2</td>
<td>222.71</td>
<td>1.36E-06</td>
<td>0.4216</td>
<td>0.8080</td>
<td>0.5095</td>
<td>n/a</td>
</tr>
<tr>
<td>rs2726648</td>
<td>KIT</td>
<td>4</td>
<td>55.53</td>
<td>4.58E-06</td>
<td>0.1391</td>
<td>0.5155</td>
<td>0.7350</td>
<td>n/a</td>
</tr>
<tr>
<td>rs16891982*</td>
<td>MATP</td>
<td>5</td>
<td>33.99</td>
<td>2.20E-16</td>
<td>2.08E-05</td>
<td>0.0985</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>rs2333857</td>
<td>OPRM1</td>
<td>6</td>
<td>154.76</td>
<td>1.74E-07</td>
<td>0.9854</td>
<td>0.1175</td>
<td>0.0197</td>
<td>0.0700</td>
</tr>
<tr>
<td>rs6917661</td>
<td>OPRM1</td>
<td>6</td>
<td>154.81</td>
<td>1.82E-08</td>
<td>0.1837</td>
<td>0.2588</td>
<td>0.2627</td>
<td></td>
</tr>
<tr>
<td>rs12668421</td>
<td>EGFR</td>
<td>7</td>
<td>54.88</td>
<td>1.36E-12</td>
<td>0.0661</td>
<td>0.7150</td>
<td>0.9492</td>
<td></td>
</tr>
<tr>
<td>rs11238349*</td>
<td>EGFR</td>
<td>7</td>
<td>54.93</td>
<td>4.57E-14</td>
<td>0.3415</td>
<td>0.5865</td>
<td>0.1430</td>
<td>0.0061</td>
</tr>
<tr>
<td>rs4948023</td>
<td>EGFR</td>
<td>7</td>
<td>55.23</td>
<td>9.15E-07</td>
<td>0.8851</td>
<td>0.1231</td>
<td>0.0168</td>
<td></td>
</tr>
<tr>
<td>rs10736126</td>
<td>HPS1</td>
<td>10</td>
<td>100.02</td>
<td>1.20E-08</td>
<td>0.4281</td>
<td>0.9002</td>
<td>0.4619</td>
<td>n/a</td>
</tr>
<tr>
<td>rs705698</td>
<td>SILV</td>
<td>12</td>
<td>54.67</td>
<td>9.19E-04</td>
<td>0.2013</td>
<td>0.5248</td>
<td>0.7586</td>
<td>n/a</td>
</tr>
<tr>
<td>rs12439639*</td>
<td>PLDN</td>
<td>15</td>
<td>43.63</td>
<td>2.20E-16</td>
<td>0.0509</td>
<td>0.1453</td>
<td>0.0176</td>
<td>n/a</td>
</tr>
<tr>
<td>rs1426654*</td>
<td>SLC24A5</td>
<td>15</td>
<td>46.21</td>
<td>2.20E-16</td>
<td>1.63E-05</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>rs12396</td>
<td>MYO5A</td>
<td>15</td>
<td>50.20</td>
<td>2.20E-16</td>
<td>0.0793</td>
<td>0.2158</td>
<td>0.0295</td>
<td></td>
</tr>
<tr>
<td>rs8026828</td>
<td>MYO5A</td>
<td>15</td>
<td>50.25</td>
<td>2.20E-16</td>
<td>0.1406</td>
<td>0.2353</td>
<td>0.0254</td>
<td>0.1288</td>
</tr>
<tr>
<td>rs4776017</td>
<td>MYO5A</td>
<td>15</td>
<td>50.29</td>
<td>4.96E-12</td>
<td>0.2708</td>
<td>0.3164</td>
<td>0.0601</td>
<td></td>
</tr>
<tr>
<td>rs11632529</td>
<td>RAB27</td>
<td>15</td>
<td>53.12</td>
<td>1.50E-03</td>
<td>0.5748</td>
<td>0.5723</td>
<td>0.2830</td>
<td></td>
</tr>
<tr>
<td>rs16976220</td>
<td>RAB27</td>
<td>15</td>
<td>53.33</td>
<td>1.28E-08</td>
<td>0.0418</td>
<td>0.4543</td>
<td>0.5211</td>
<td>0.2156</td>
</tr>
<tr>
<td>rs472579</td>
<td>RAB27</td>
<td>15</td>
<td>53.46</td>
<td>2.44E-13</td>
<td>0.5363</td>
<td>0.2751</td>
<td>0.0818</td>
<td></td>
</tr>
<tr>
<td>rs8105923</td>
<td>AP3D1</td>
<td>19</td>
<td>1.86</td>
<td>0.11480</td>
<td>0.4423</td>
<td>0.0562</td>
<td>0.0223</td>
<td>0.0754</td>
</tr>
<tr>
<td>rs4807227</td>
<td>AP3D1</td>
<td>19</td>
<td>2.27</td>
<td>0.57960</td>
<td>0.4124</td>
<td>0.2918</td>
<td>0.2691</td>
<td></td>
</tr>
<tr>
<td>rs562926</td>
<td>ATRN</td>
<td>20</td>
<td>3.57</td>
<td>2.96E-04</td>
<td>0.7248</td>
<td>0.3434</td>
<td>0.0913</td>
<td></td>
</tr>
<tr>
<td>rs3859664</td>
<td>ATRN</td>
<td>20</td>
<td>3.62</td>
<td>3.25E-03</td>
<td>0.8368</td>
<td>0.2246</td>
<td>0.1226</td>
<td></td>
</tr>
<tr>
<td>rs1015361*</td>
<td>ASIP</td>
<td>20</td>
<td>32.20</td>
<td>0.22100</td>
<td>0.2627</td>
<td>0.0342</td>
<td>0.0229</td>
<td>0.0002</td>
</tr>
<tr>
<td>rs6087577</td>
<td>ASIP</td>
<td>20</td>
<td>32.42</td>
<td>1.17E-03</td>
<td>0.2491</td>
<td>0.0037</td>
<td>0.0005</td>
<td></td>
</tr>
</tbody>
</table>
Because *SLC24A5* and *MATP* have such a large influence on skin pigmentation, it is possible that their effect is masking that of another gene with a smaller effect. To determine if this is the case, the genotypes at these loci were included as covariates. Noting that after the inclusion of *SLC24A5* as a covariate, *MATP* was no longer significantly contributing to the variance in skin pigmentation, further consideration should be given both to the analyses including and not including this genotype as a covariate. Additionally, for genes including more than one SNP, haplotypes were calculated using the program fastPHASE [38]. Because the causative SNPs are likely not one of the SNPs being tested in this research, but could be near the test SNPs, creating haplotypes can increase the likelihood of finding an association by more accurately capturing the variation in the gene.

These results are incompletely overlapping with those from ADMIXMAP. The strongest association based the Kruskal-Wallis results is *ASIP* for which the *p*-value decreased as more of the covariates were included in the analysis. This is an indication that this may be a real result because the decreasing *p*-value indicates that as the covariates account for more and more of the variation in melanin, a larger proportion of the remaining variation may be accounted for by *ASIP*. *ATRN* and *AP3D1*, which also appeared to be associated with skin pigmentation using ADMIXMAP, show little evidence of an association in these results. Finally, *PAX3*, which showed a strong signal in ADMIXMAP, does not have a single significant or suggestive result in these analyses. However, *EGFR* also shows a suggestive, although not significant *p*-value in the haplotype analysis. A low z-score in ADMIXMAP could be an indication that there is no effect of this gene on skin pigmentation or it may indicate that this result is spurious. In either case, if the effect of the gene is weak, or the causative SNP is far from the tests SNPs, additional genotyping to increase sample size should resolve the question.
Mann-Whitney $U$ Tests

To determine if these associations are indicating actual mean differences in the distributions of melanin and what the magnitudes of these distributions are, Mann-Whitney $U$ (two-sample Wilcoxon) tests were used to compare the mean melanin values for each genotype. To account for the influence of ancestry, $SLC24A5$, and $MATP$ on the total variation in $M$, a series of regressions was performed, first $M$–ancestry with the residual values regressed onto $SLC24A5$ and $MATP$ in turn. The residual values, referred to as adjusted $M$, were used in the Mann-Whitney tests and are graphed as histograms in Figure 4-11.
As expected, SLC24A5 and MATP were clearly significant for mean differences. The effect appears to additive with significant differences for both genes between the heterozygote and both homozygotes. The magnitude of the difference is about 3 adjusted melanin units. Unexpectedly, however, both ASIP(rs1015361) and PAX3 also showed significant p-values at $\alpha = 0.5$. When ASIP was treated as a dominant allele with G dominant to A (more common in Indigenous Americans), individuals who were AA were approximately 1 adjusted melanin unit darker than individuals who were AG or GG with a $p$ value of 0.0045. PAX3 shows the same apparent pattern of dominance with G dominant to A (more common in Indigenous Americans) resulting in an even larger melanin difference of 1.4 adjusted melanin units with AA individuals more darkly pigmented ($p = 0.0101$).

**Conclusions**

This sample was severely limited by the need to subdivide it into the four constituent populations for some analyses. As a result, the statistical power of the sample means that genes with a smaller effect on skin pigmentation could be missed. Nevertheless, a number of genes show evidence of admixture linkage with skin pigmentation differences between Indigenous American and European populations, but PAX3 and ASIP show the strongest results. Although neither of these plays as significant a role as SLC24A5 and MATP, they each have between one-third and one-half of the effect size of these genes. This is a remarkable result as it the first time that these genes have been linked to darker skin pigmentation in Indigenous Americans.
References


Chapter 5

Conclusion

The three goals of this study were to determine the distribution of ancestry among three ethnic groups from highland Colombia, examine evidence of selection at pigmentation genes in New World populations for the first time, and use admixture linkage analysis to determine the role that these selection-nominated candidate genes play in determining skin pigmentation differences between European and Indigenous American populations. The results of these three chapters can be distilled into a few main points that are worth reviewing as the implications may extend to other, future work.

Relationship between Biogeographic Ancestry and Ethnicity

Colombia is among the most diverse countries in Central or South America [1] and is of particular interest to anthropologists because of its location at the base of the isthmus that is the only land route for Homo sapiens to have used to colonize South America. Comparison of the three Colombian ethnic groups – Campesino/Indigena, Mestizo, and Afro-Colombian – in terms of biogeographic ancestry (BGA) based on both non-recombinant Y (NRY) and autosomal ancestry informative markers (AIMs) resulted in some interesting conclusions. In general, the distribution of both categories of markers revealed patterns of genetic variation consistent with prior expectations. The individuals who self-identify as Campesino or Indigena are the most likely to have an Indigenous American Y haplotype or a high proportion of Indigenous American ancestry as estimated from the autosomal markers. Similarly, Mestizo individuals are expected to have a large proportion of European ancestry and Afro-Colombians are expected to show substantial ancestry from Africa and in both cases the results conform to these expectations. The
analysis shows statistically significant differences in ancestry among these ethnic groups. However, this is not the full story. The NRY and autosomal AIMs results both showed that these groups are far from homogenous with distributions of BGA that overlap substantially. Clearly, BGA is not the only determinant of self-identified ethnicity among these individuals.

Ethnic identity is shaped by a myriad of factors including presumed ancestry (most individuals are not basing their identity on the results of a genetic test). Some of these traits may be social including the ethnic designations of friends and relatives, socio-economic status, nativity, residence, occupation, language, political affiliations, etc. Some other components of identity are biological and can include skin, hair, and iris pigmentation, hair texture, build, facial features, and more. The interplay of these complex and frequently dynamic factors is difficult to test, but the way that individuals create a self-identity is an interesting question. Skin pigmentation, because it is a clearly visible and difficult to alter phenotype, is an attractive candidate for a marker of ethnicity and self-conception. In the Colombian sample, the distribution of ancestry and skin pigmentation among the three ethnic groups supported a hypothesis that individuals with similar proportional ancestry levels may be relying, in part, on their skin pigmentation in defining their membership in different ethnic groups. In particular, a number of Afro-Colombian individuals have small proportions of West African ancestry but fairly dark skin. However, the attempt to use the available data to determine the influence of skin pigmentation on ethnic self-identification was inconclusive. This is due in part to the fact that skin pigmentation, like many other phenotypes frequently associated with racial or ethnic distinctions, is imperfectly correlated with BGA. As a result, it is difficult to separate the influence of these two factors in on self-reported ethnicity. Future research in this region could benefit from a broader collection of both the biological phenotypes and social factors that may be influencing ethnic identity so that self-conception in Colombia can be better understood.
Genes Contributing to Skin Pigmentation Variation

This work was the first to look for signatures of selection at skin pigmentation genes in New World populations. Using three tests for selection in two separate sets of sample populations, fourteen genes were identified as potentially having undergone selection out of an original list of 72 candidate genes. SNPs showing high allele frequency differences (δ) between Indigenous American and European populations were genotyped in four admixed populations from the southern United States, Mexico, and Colombia. Using admixture linkage analysis, two genes, *PAX3* and *ASIP* stood out as having derived alleles that were associated with darker skin pigmentation.

PAX3

*PAX3* (Paired Box 3, homolog of Pax3 in mice) is a transcription factor involved in a number of developmental pathways including those for the central nervous system, skeletal muscles, and neural-crest derived cell types. It is also integral to melanocyte proliferation, migration, and differentiation where it is has been shown in mouse models to interact with other pigmentation genes including *TYR*, *DCT*, and *MITF*. These different roles are fulfilled by virtue of extensive alternate splicing. [2] Although its role in skin pigmentation has been well established, previous variation has only been linked to pathogenic conditions, primarily Waardenburg Syndrome. Waardenburg Syndrome is an auditory-pigmentary condition characterized by congenital hearing loss and pigmentation abnormalities including white forelocks and premature graying of the hair, heterochromia or bright blue irises. [3] This report is the first evidence of normal variation linked to *PAX3*. 
PAX3 shows clear evidence of selection in Indigenous Americans with patterns of significance in the same region of the gene (chromosome 2 at 222.65 Mbp) on all three tests and in both sets of population samples (Figure 2-6). This pattern was among the strongest seen among all of the selection-nominated candidate genes. The significant z-score in the ADMIXMAP analysis demonstrated that the gene must have an effect size of at least 1.2 to have been detected in that analysis. Additionally, after adjustment for genome-wide BGA and for the effects of SLC24A5 and MATP, PAX3 still influenced skin pigmentation as seen in the mean differences in adjusted melanin index values between the AA and AA/AG genotypes.

ASIP

Like PAX3, ASIP is a gene known to be involved in melanogenesis. The agouti signaling protein (homolog of the mouse agouti gene) blocks the binding of α-melanocyte stimulating hormone (α-MSH) to MC1R which results in lower tyrosinase activity and a decrease in melanocyte proliferation. [4] ASIP has been previously associated with variation in human skin pigmentation in African Americans [5], as well as with red hair, freckling, and melanoma in European populations [6-8].

In the analysis of selection, ASIP shows negative, but not significant lnRH values across the gene and positive, but not significant, values for LSBL. However, the Tajima’s D and Tajima’s D difference results were among the strongest of any of the selection-nominated candidate genes. In the ADMIXMAP results, only the San Luis Valley sample showed a significant z-score for allelic association. This is likely because the effect size, shown to be approximately one-third that of SLC24A5 and MATP, was at the cusp of what could be detected with these analyses. The haplotype of the three ASIP SNPs genotyped in this study was, however, the only one to show statistically significant association with variation in skin
pigmentation after Bonferroni corrections for multiple testing. Future research into PAX3 and ASIP should include additional genotyping or sequencing to localize the functional mutations in these genes. Additionally, other admixed populations should be genotyped to confirm these results.

SLC24A5 and MATP

This research also demonstrated a role for SLC24A5 and MATP (SLC45A2) in skin pigmentation differences between Indigenous American and European populations. These genes showed by far the strongest association with skin pigmentation in ADMIXMAP and ANOVA analyses with each having roughly double the effect size of ASIP or PAX3. However, these genes are not associated with episodes of skin darkening in Indigenous American populations, but with the evolution of lighter skin pigmentation in Europeans as has been previously shown [9-13].

Necessity of Testing Functionality for Signatures of Selection

Beyond discussing the selection-nominated candidate genes that do appear to contribute to variation between Indigenous American and European skin pigmentation, it is important to acknowledge that many genes that showed evidence of selection are not associated with skin color variation. There are many possible explanations for these apparent false positives in the selection results including both reasons why neutrally evolving alleles may appear to have undergone selection and why pigmentation candidate genes that have undergone selection in Indigenous American populations might not influence pigmentation.

When assessing signatures of selection, one must determine an $\alpha$ significance cutoff. Selecting an accurate cutoff is difficult without a priori knowledge of the proportion of genomic
regions undergoing selection. When the most extreme 5% of regions are considered to have undergone selection, some neutrally evolving loci are going to be included unless exactly 5% of regions have, in fact, undergone selection. Factors such as population sub-structure can also lead to false-positive results for tests of selection [14] as can genetic drift. By chance, certain alleles may drift to fixation, creating the illusion of a signature of selection in the genome. This stochastic effect is especially likely in populations like the Indigenous Americans who have experienced at least one, if not multiple, bottlenecks. During a bottleneck, some genetic variation is lost and as the population recovers from the bottleneck, wide haplotypes created by the loss of variation may increase in frequency simulating the appearance of selection. The use of the empirical distribution should offset this effect, but there may still be some residual impact of the bottlenecks.

Neutrally evolving genes that appear to have undergone selection are not the only explanation for a lack of association. Many of the selection-nominated pigmentation genes show clear evidence of pleiotropy. Because they are involved in biological pathways other than melanogenesis, these genes may contain mutations that are adaptive in the New World but do not influence skin color. Selection may have favored these mutations without any impact on skin pigmentation.

Finally, there is the issue of statistical power. In this study, only genes with an effect size of approximately 1.2 could be identified. Genes with small effects, as might be common for complex traits like skin pigmentation, might be missed. If this is the case, the results of the selection screen were not false-positives; instead, the results of the admixture linkage analysis were false negatives. Whatever the cause, this research demonstrates that identification of selection at a particular gene is a good first step in identifying functional variation, but additional research is needed to validate these results.
Future Directions

Genes for Future Investigation

Three selection nominated candidate genes – PLDN, MYO5A, and RAB27A – located less than five megabases from SLC24A5 showed suggestive evidence for association with variation in skin pigmentation, but the results were inconclusive. Due to the long regions of linkage disequilibrium generated by admixture, there is a possibility that these genes are showing evidence of association with skin pigmentation because the European alleles at each SNP are in strong LD with the causative SNP in SLC24A5. However, several lines of evidence suggest that these genes may have an independent effect. First, each of these genes show evidence of selection in Indigenous American populations and SLC24A5 does not (see Appendix II). Second, MYO5A and RAB27A appear to show peaks for allelic association in the ADMIXMAP analyses that are independent from the peak associated with SLC24A5. Third, MYO5A and PLDN have results on the Kruskal-Wallis analysis of variance that are suggestive (not significant after Bonferroni correction) after including ancestry and SLC24A5 and MATP genotypes as covariates. This decrease in $p$-value may indicate that as the proportion of variation in skin pigmentation caused by those genes is accounted for, the correlation between skin pigmentation and MYO5A and PLDN increases. None of these genes, however, is associated with a mean difference in melanin among the genotypes after adjusting for ancestry and the genotypes at SLC24A5 and MATP.

To determine if PLDN, MYO5A, and/or RAB27A play a role in skin pigmentation variation, the sample size and the marker density will both need to be increased. Increasing the sample size will increase the likelihood of identifying individuals in whom recombination has occurred along this region as well as increasing statistical power to capture smaller effect sizes.
To do this, additional admixed individuals from previously collected samples of individuals from Brazil and the U.S. will be genotyped. Increasing resolution will also allow for the detection of recombinant haplotypes and, by genotyping more loci, will narrow down the region containing the causative allele. This will be achieved either by genotyping additional SNPs in the region or by sequencing.

**Dating Rise of Mutations**

Having identified *ASIP* and *PAX3* as contributing to variation in skin pigmentation between Indigenous Americans and Europeans, an interesting next step would be to date the derived mutations using short tandem repeat markers (STRs). Dating these mutations is important for understanding the history of the mutations and ultimately the evolution of skin pigmentation world-wide. Identifying the approximate date when and likely populations where the derived mutations increased in frequency will allow us to evaluate hypotheses including the likely skin color of the populations that migrated into the New World. Microsatellite data has been successfully used to reconstruct demographic expansions [15-17] and to date specific mutational events [18-20]. They have recently been applied to four pigmentation genes contributing to variation in skin pigmentation lightening in European populations (S. Beleza et al., in preparation).

The STRs can be typed in the same admixed populations used in this analysis which will allow for the combined haplotypes formed by the SNPs and the STRs to be analyzed. The relationships between these haplotypes will be visualized using NETWORK 4.5 [21] which can construct networks using the reduced median [22] and median-joining [23] methods described in detail elsewhere. These networks indicate the genetic distance between the haplotypes, but coalescence and/or forward simulations are necessary to estimate the time of the mutation.
Coalescence simulations could be performed using SIMCOAL 2.1.2 software [24] which uses a coalescent backward approach to simulate mutations starting from the most recent common ancestor (MRCA) and adding mutations assuming a constant mutation rate [25, 26]. Several permutations of the demographic history of the Indigenous American populations will have to be considered because of the conflicting hypotheses regarding the number, size, and timing of migrations into the Americans [27-29]. Additionally, forward simulations can be done using a Monte Carlo method which also incorporates drift, selection, recombination and mutation to estimate time to most recent common ancestor [20]. The advantage of the forward simulation is that it allows an estimation of the strength of selection necessary to produce the observed variation.

Both coalescent models are based on comparing the amount of variation present in the haplotypes defined by the tested markers. If the derived allele is at high frequency in the population and there is substantial variation at the linked STR loci, this would support a neutral model of an older SNP mutation which has accumulated new STR mutations and recombination events over time. However, if this same high-frequency allele defines haplotypes with very variation in the STR loci, this supports a model of selection in which a single haplotype (or perhaps a few related haplotypes) have rapidly increased in frequency, not allowing sufficient time for mutations to occur in the STRs or for recombination to break down the haplotype block. Dating the rise of the mutations in ASIP and PAX3 will complete the story of the evolution of these genes and their contribution to darker skin pigmentation in Indigenous American populations.

As the first report of signatures of selection in the Americas and the first identification of ASIP and PAX3 as genes contributing to darker skin pigmentation among Indigenous Americans, this research contributes significantly to a previously understudied area of research. There are
many more questions to be answered regarding the evolution of skin pigmentation in the New World, but this comprises an important first step.


Appendix A

PERL Scripts for Calculating Test Statistics

LSBL

open(ofile2,">lsbl.txt");
print ofile2
"rs\tchr\tpos\tfstEuropean_EastAsian\tfstEastAsian_NativeAmerican\tfstEuropean_NativeAmerican\nfstEuropean\nEastAsian\nNativeAmerican\ntypingresults\n"; # change to match populations with fst 1-2, 2-3, 1-3.

for($x=1;$x<=22;$x++) {
    $foldername="chr".$x;
    $path="./".$foldername;
    print "$path\n";
    opendir(DIR, $path);
    @files=sort(grep(/$/,readdir(DIR)));
    closedir(DIR);

    @freq1=[];
    @freq2=[];
    @freq3=[];
    @count1=[];
    @count2=[];
    @count3=[];
    @rs=[];
    @chr=[];
    @pos=[];

    foreach $filename (@files) {
        if($filename=~/txt/ && ($filename=~/Europe/|| $filename=~/EastAsia/ || $filename=~/America/) ) {  #change to match populations 1,2,3
            print("$filename\n") unless -d;
            $file=$path.".".$filename;
            print "$file\n";

            open(ifile,"$file");
            $j=0;
            while(eof(ifile)==0) {
                $line=<ifile>;
                chop($line);
                @array=split("\t",$line);
            }
if($array[4]==0) {
    $rs[$j]=$array[0];
    $chr[$j]=$array[1];
    $pos[$j]=$array[2];
    $freq=$array[3];
    $count=$array[4];
} else {
    $rs[$j]=$array[0];
    $chr[$j]=$array[1];
    $pos[$j]=$array[2];
    $count="NA";
    $freq="NA";
}

change to pop1
if($filename=~/Europe/) {
    #
    $freq1[$j]=$freq;
    $count1[$j]=$count;
    $j++;
}

#change to pop2
if($filename=~/EastAsia/) {
    #
    $freq2[$j]=$freq;
    $count2[$j]=$count;
    $j++;
}

#change to pop3
if($filename=~/America/) {
    #
    $freq3[$j]=$freq;
    $count3[$j]=$count;
    $j++;
}

} close(ifile);

$filename1="chr",".x",".txt";
open(ofile1,">$filename1");
print ofile1 "rs\chr\pos\EuropeLSBL\tEastAsiaSBL\tAmericaLSBL\ttypingresults\n"; # change
to match pops, same order
for($a=0;$a<#freq1;$a++) {
    $truecall=$count1[$a].",".$count2[$a].",".$count3[$a];
    $hs=2*$count1[$a]/(2*$count1[$a]-1)*$freq1[$a]*(1-$freq1[$a])+2*$count2[$a]/(2*$count2[$a]-1)*$freq2[$a]*(1-$freq2[$a]);
    $ht=0.5*($hs+$freq1[$a]*(1-$freq2[$a])+$freq2[$a]*(1-$freq1[$a]));
    if($ht==0 & & $hs==0) {
        $FST12="NA";
    } else {
        $FST12=2*($ht-$hs)/(2*$ht-$hs);
    }
    $hs=2*$count3[$a]/(2*$count3[$a]-1)*$freq3[$a]*(1-$freq3[$a])+2*$count2[$a]/(2*$count2[$a]-1)*$freq2[$a]*(1-$freq2[$a]);
    $ht=0.5*($hs+$freq3[$a]*(1-$freq2[$a])+$freq2[$a]*(1-$freq3[$a]));
    if($ht==0 & & $hs==0) {
        $FST23="NA";
    } else {
        $FST23=2*($ht-$hs)/(2*$ht-$hs);
    }
    $hs=2*$count1[$a]/(2*$count1[$a]-1)*$freq1[$a]*(1-$freq1[$a])+2*$count3[$a]/(2*$count3[$a]-1)*$freq3[$a]*(1-$freq3[$a]);
    $ht=0.5*($hs+$freq1[$a]*(1-$freq3[$a])+$freq3[$a]*(1-$freq1[$a]));
    if($ht==0 & & $hs==0) {
        $FST13="NA";
    } else {
        $FST13=2*($ht-$hs)/(2*$ht-$hs);
    }
    if(($FST12)eq"NA" || ($FST23)eq"NA" || ($FST13)eq"NA") {
        $bl1="NA";
        $bl2="NA";
        $bl3="NA";
    } else {
        $bl1=($FST12+$FST13-$FST23)/2;
        $bl2=($FST12+$FST23-$FST13)/2;
        $bl3=($FST23+$FST13-$FST12)/2;
    }
    print ofile1 "$rs[$a]\t$chr[$a]\t$pos[$a]\t$bl1\t$bl2\t$bl3\t$truecall\n";
    print ofile2 "$rs[$a]\t$chr[$a]\t$pos[$a]\t$bl1\t$bl2\t$bl3\t$truecall\n";
} close(ofile1);
} close(ofile2);
# initialize hashes
    my %chr=();
    my %pos=();
    my %ceu=();
    my %na=();
    my %yri=();
    my %ea=();
    my @markers=();

# for each file in each folder, add to hash

for($x=1;$x<=22;$x++) {
    $foldername="chr".$x;
    $path="./$foldername; print "$path
";
    opendir(DIR, $path);
    @files=sort(grep(/$/,.readdir(DIR)));
    closedir(DIR);

    foreach $filename (@files) {
        if(($filename=~/txt/) && ($filename=~/EuropeanAm/) ||
           ($filename=~/NativeAm/) ||($filename=~/YRI/) || ($filename=~/EastAsian/)) {
            print("$filename
") unless
            $file=$path."/".$filename;
            #print "$file
";
            open(ifile,"$file");

            while(eof(ifile)==0) {
                $line=<ifile>;
                chomp($line);
                @array=split("\t",$line);
                $id=$array[0];

                $freq1=$array[3];
                $exphet=1-$freq1*$freq1-(1-$freq1)*(1-$freq1);  #note that this

                is changed from Mao's original script

                # if not already defined, add to hash
                if (($id=~SNP/) && (!exists $chr{$id})){
                    $chr{$id}=$array[1];
                }
            }
        }
    }
$pos{Sid}=array[2];
push (@markers, Sid);
}

# put into hashes, 4 decimal expected heterozygosities
if($filename=~/EuropeanAm/) {
    $ceu{Sid}=substr($exphet,0,6);
} if($filename=~/NativeAm/) {
    $na{Sid}=substr($exphet,0,6);
} if($filename=~/YRI/) {
    $yri{Sid}=substr($exphet,0,6);
} if($filename=~/EastAsian/) {
    $ea{Sid}=substr($exphet,0,6);
}
close (ifile);

open(ofile1,">Expected_Het.txt");
print ofile1 "rsID
tchr
tpos
tExpHetCEU
tExpHetYRI
tExpHetEA
tExpHetNA
";
open (ofile2, ">LnRH_Results.txt");
print ofile2 "rsID
tchr
tpos
tlnRH(NativeAmerican/EastAsian)
tlnRH(NativeAmerican/European)
tlnRH(NativeAmerican/African)
tlnRH(EastAsian/European)
tlnRH(EastAsian/African)
tlnRH(European/African)";

foreach $SNPid(@markers) {
    print ofile1 $SNPid.".t".Schr{SNPid}.".t".Spos{SNPid}.".t".Sceu{SNPid}.".t".Syri{SNPid}.".t".Sea{SNPid}.".t".Sna{SNPid}.".n";
    if ($ceu{SNPid} == 0) {
        $ceu{SNPid}=0.0001;  # for cases where expected heterozygosity is 0 (i.e. there is fixation), the E(H) value is being changed to 0.0001
    } if ($yri{SNPid} == 0) {
        $yri{SNPid}=0.0001;
    } if ($ea{SNPid} == 0) {
        $ea{SNPid}=0.0001;
    }
if ($na{$SNPid} == 0) {
    $na{$SNPid}=0.0001;
}

$comp_ceu=(1/(1-$ceu{$SNPid}))**2-1;  #creates top or bottom half of calculation
$comp_yri=(1/(1-$yri{$SNPid}))**2-1;
$comp_ea=(1/(1-$ea{$SNPid}))**2-1;
$comp_na=(1/(1-$na{$SNPid}))**2-1;

$lnRH1=log($comp_na/$comp_ea);
$lnRH2=log($comp_na/$comp_ceu);
$lnRH3=log($comp_na/$comp_yri);
$lnRH4=log($comp_ea/$comp_ceu);
$lnRH5=log($comp_ea/$comp_yri);
$lnRH6=log($comp_ceu/$comp_yri);

print ofile2
SSNPid."\t".$chr{$SNPid}."\t".Spos{$SNPid}."\t".lnRH1."\t".lnRH2."\t".lnRH3."\t".lnRH4."\t".lnRH5."\t".lnRH6."\t"
}
close (ofile1);
close (ofile2);

Tajima’s D

open(ifile,"par.txt");

$line=<ifile>;
chop($line);

$incr=$line;
$line=<ifile>;
chop($line);
$no=$line;

for($i=0;$i<$no;$i++) {
    $line=<ifile>;
    chop($line);
    $file[$i]=$line;
}
close(ifile);
for($i=0;$i<$no;$i++) {
    $filesuf=$file[$i];
    for($j=1;$j<=22;$j++) {
        $ifile="chr".$j."_".$filesuf;
        print "$ifile\n"
        $reffile="chr".$j."_ref.txt";
        open(reffile,$reffile) || die "Can't open references file: $reffile\n";
        open(ifile,$ifile) || die "Can't open input file: $ifile\n";
        while(eof(reffile)==0) {
            $line=<reffile>
            if($line=~/^P/) {
                chop($line);
                @pos=split(" ",@line);
                $tot=$#pos-1;
                last;
            }
        }
        $n=0;
        for($i=0;$i<=$tot;$i++) {
            $good[$i]=0;
        }
        while(eof(ifile)==0) {
            $line=<ifile>
            if($line=~/^#/) {
                $line=<ifile>
                chop($line);
                my @array=split(" ",@line);
                for($i=0;$i<=$tot;$i++) {
                    if($array[$i]eq"A" ||$array[$i]eq"B" ) {
                        $good[$i]++;
                    }
                }$data[$n]=(@array);
                $n++;
                $line=<ifile>
                chop($line);
                my @array=split(" ",@line);
                for($i=0;$i<=$tot;$i++) {
                    if($array[$i]eq"A" ||$array[$i]eq"B" ) {
                        $good[$i]++;
                    }
                }$data[$n]=(@array);
                $n++;
                $line=<ifile>
                chop($line);
                my @array=split(" ",@line);
                for($i=0;$i<=$tot;$i++) {
                    if($array[$i]eq"A" ||$array[$i]eq"B" ) {
                        $good[$i]++;
                    }
                }$data[$n]=(@array);
                $n++;
            }
        }
    }
}
$data[$n]=(@array;
$n++;  
}
}
$data[$n]=(@array;
$n++;  
}
}
if($n==0) {
} else {
    print "$tot\n";
    $ofile="Taj_".$ifile;
    open(ofile,">$ofile");
    $ca=0;
    @cpos=();
    for($a=0;$a<=$tot;$a++) {
        if($good[$a]>($n/2) ){
            for($b=0;$b<$n;$b++) {
                $seq=$data[$b];
                if($seq[$a]eq"A") {
                    $cseq[$ca][$b]=1;
                }
                if($seq[$a]eq"B") {
                    $cseq[$ca][$b]=2;
                }
                if($seq[$a]eq"?" ) {
                    $cseq[$ca][$b]=-1;
                }
                }
                $cpos[$ca]=$pos[$a+1];
                $ca++;
            }
        }
    print $#cpos, "\n";
    $start=0;
    for($int=$cpos[$start];$int<=($cpos[$start]+$incr);$int=$int+$incr/4) {
        for($i=0;$i<=$#cpos;$i++){
            if($cpos[$i]>$int ) {
                $cursor=$i;
                last;
            }
        }
    }
}
for($inti=$int;$cursor<=$#cpos;$inti=$inti+$incr) {
    $size=0;
    while($cpos[$cursor]>$inti+$incr &&
    $cpos[$cursor]>$inti) {
        for($b=0;$b<$n;$b++) {
            $cseq[$b] = $cseq[$cursor][$b];
            $size++;
        }
    }
    if($size>0) {
        $s = taj_s($size,$n);
        $pi = taj_pi($size,$n);
        $v = taj_var($s,$n);
        $a1 = 0;
        for($a=$n-1;$a>=1;$a--) {
            $a1 = $a1+1/$a;
        }
        if($v==0) {
            $tajimaD=0;
        } else {
            $tajimaD=($pi-$s/$a1)/$v;
        }
        print ofile "$inti\t$tajimaD\t*size\n";
    }
}
}
}
}
}
close(ofile);
close(ifile);
close(refile);

##S

sub taj_s ($$) {
    my ($size,$n)=@_; 
    my $s=0;
    for(my $i=0;$i<$size;$i++) {
        my $A=0;
        my $B=0;
        for(my $j=0; $j<$n;$j++) {
            if($cseq[$i][$j]==1) {
                $A++;
            } else {
                $B++;
            }
        }
    }
}
my $freq=$A/($B+$A);
if($freq==1 || $freq ==0) {
    $s++;
} else {
}
return($s);

## pi

sub taj_pi ($$) {
    my ($size,$n)=@_; 
    my $pi=0; 
    my $pisum=0; 
    for(my $i=0;$i<($n-1);$i++) {
        for(my $j=$i+1:$j<$n;$j++) {
            for (my $k=0;$k<$size;$k++) {
                if($curseq[$k][$j]>0 && $curseq[$k][$i]>0 ){
                    if($curseq[$k][$j]==$curseq[$k][$i]) {
                        $pisum++;
                    } else {
                        
                    }
                }
            }
        }
    } 
    $pi=$pisum*2/($n-1)/$n;
    return($pi);
}

##variance

sub taj_var ($$) {
    my ($s, $n)=@_; 
    my $a1=0; 
    for(my $a=$n-1;$a>=1;$a--) {
        $a1=$a1+1/$a;
    }
    my $a2=0; 
    for(my $a=$n-1;$a>=1;$a--) {
        $a2=$a2+1/$a/$a;
    }
    my $b1=($n+1)/3/($n-1);
    my $b2=2*($n*$n+$n+3)/9/$n/($n-1);
    my $c1=$b1-1/$a1;
my $c2=$b2-($n+2)/$a1/$n+$a2/$a1/$a1;
my $var=sqrt($c1/$a1*$s+$c2/($a1*$a1+$a2)*$s*(s-1));

Tajima’s D Difference

#initialize
$chr = 0;
$pos = 0;
$taj = 0;
$SNPs = 0;
$chr2 = 0;
$pos2 = 0;
$taj2 = 0;
$SNPs2 = 0;
$taj_diff = 0;
$x = 0;
$y=0;

open (ifile1, "Tajima_NativeAmerican.txt");
open (ifile2, "Tajima_EastAsian.txt");
open (ofile, ">Tajima_Diff_NA-EA.txt");

print ofile "chr\tposition\tTaj_Diff_NA-EA\tSNPs\n";

while (eof(ifile1) == 0){
    $fileNA = <ifile1>;
    chomp $fileNA;
    ($chr, $pos, $taj, $SNPs) = split(\t, $fileNA);

    $fileEA = <ifile2>;
    chomp $fileEA;
    ($chr2, $pos2, $taj2, $SNPs2) = split(\t, $fileEA);

    $x++;
    if (($chr==$chr2) && ($pos==$pos2)) {
        $taj_diff = $taj-$taj2;
        print ofile $chr.\t$pos.\t$taj_diff.\t$SNPs.\t$SNPs2.\n;
        $y++;
    }
}
print "$x windows\n$y lines\n"

close ifile1;
close ifile2;
close ofile;

exit;

Normalization of Tajima's D

#declare and initialize
$chr = "";
$position = "";
$tdea = "";
$snps = "";
$normaldea = "";

.#have superloop for chromosome#

$filename1="normalized_Tajima_NA-EA.txt";

#open an outfile for writing
open(OFILE, ">$filename1");

#print to OFILE
print OFILE "chr\tposition\tTajima_D_NA-EA\tSNPS\tNormalized_D_NA-EA\n";

#open each chromosome file
open(INFILE, "Tajima_Diff_NA-EA.txt");

while(eof(INFILE)==0) {
  $line = <INFILE>
  chomp($line);

  #split each line into an array
  @tajimasd = split("\t",$line);
  $chr = $tajimasd[0];
  $position = $tajimasd[1];
  $tdea = $tajimasd[2];
  $snps = $tajimasd[3];
  $normaldea = (($tdea-0.50292)/1.22112); #must change mean and stdev for each
  #print to OFILE
  unless ($position eq 'position') {
    print OFILE "$chr\t$position\t$tdea\t$snps\t$normaldea\n";
  }

}

}
close(INFILE);
close (OFILE);
exit;
Appendix B

Results of Selection Analysis

The following pages contain the results of the selection analyses lnRH, LSBL, Tajima’s D, and Tajima’s D Difference plotted for all 76 candidate genes including the regions upstream and downstream of the gene. The horizontal axis indicates the chromosomal position and the vertical axis indicates the value of the test statistic with orange bars indicating that a result in the 5% tail.
SLC7A11 LnRH HGDP

SLC7A11 LnRH Mega

SLC7A11 LSLH HGDP

SLC7A11 LSLH Mega

SLC7A11 Tajima's D Mega

SLC7A11 Tajima's D Difference Mega
# Appendix C

## Parental Population Frequencies

**BurchardAIMs Panel (Used for Colombian Sample)**

<table>
<thead>
<tr>
<th>Marker ID</th>
<th>Chr</th>
<th>cM</th>
<th>European Frequency</th>
<th>Native American Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2817611</td>
<td>1</td>
<td>27.64985</td>
<td>0.952381</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs6684063</td>
<td>1</td>
<td>59.80265</td>
<td>0.833333</td>
<td>0.166667</td>
</tr>
<tr>
<td>rs1934393</td>
<td>1</td>
<td>81.83714</td>
<td>0.842105</td>
<td>0.3</td>
</tr>
<tr>
<td>rs3768176</td>
<td>1</td>
<td>89.54529</td>
<td>0.928571</td>
<td>0.233333</td>
</tr>
<tr>
<td>rs3828121</td>
<td>1</td>
<td>116.0777</td>
<td>0.869048</td>
<td>0.3</td>
</tr>
<tr>
<td>rs3806218</td>
<td>1</td>
<td>153.7027</td>
<td>0.666667</td>
<td>0.821429</td>
</tr>
<tr>
<td>rs2592888</td>
<td>1</td>
<td>164.9981</td>
<td>0.761905</td>
<td>1</td>
</tr>
<tr>
<td>rs4657449</td>
<td>1</td>
<td>176.7128</td>
<td>0.071429</td>
<td>0.733333</td>
</tr>
<tr>
<td>rs7535375</td>
<td>1</td>
<td>268.0049</td>
<td>0.714286</td>
<td>0.8</td>
</tr>
<tr>
<td>rs1073319</td>
<td>2</td>
<td>51.22828</td>
<td>0.809524</td>
<td>0.166667</td>
</tr>
<tr>
<td>rs1470524</td>
<td>2</td>
<td>70.55513</td>
<td>0.785714</td>
<td>0.533333</td>
</tr>
<tr>
<td>rs842634</td>
<td>2</td>
<td>84.50958</td>
<td>0.738095</td>
<td>0.233333</td>
</tr>
<tr>
<td>rs10515919</td>
<td>2</td>
<td>100.3808</td>
<td>0.880952</td>
<td>0.166667</td>
</tr>
<tr>
<td>rs4852696</td>
<td>2</td>
<td>107.6342</td>
<td>0.904762</td>
<td>0.7</td>
</tr>
<tr>
<td>rs3860446</td>
<td>2</td>
<td>117.0775</td>
<td>0.357143</td>
<td>1</td>
</tr>
<tr>
<td>rs1036543</td>
<td>2</td>
<td>149.5115</td>
<td>0.02381</td>
<td>0.766667</td>
</tr>
<tr>
<td>rs2711070</td>
<td>2</td>
<td>172.0077</td>
<td>0.464286</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs868179</td>
<td>2</td>
<td>188.6715</td>
<td>0.073171</td>
<td>0</td>
</tr>
<tr>
<td>rs10497705</td>
<td>2</td>
<td>195.1086</td>
<td>0.238095</td>
<td>0.866667</td>
</tr>
<tr>
<td>rs1517634</td>
<td>2</td>
<td>229.7175</td>
<td>0.833333</td>
<td>0</td>
</tr>
<tr>
<td>rs10498255</td>
<td>2</td>
<td>240.7059</td>
<td>0.809524</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs304051</td>
<td>3</td>
<td>13.04121</td>
<td>0.02381</td>
<td>0.566667</td>
</tr>
<tr>
<td>rs1498991</td>
<td>3</td>
<td>40.80151</td>
<td>0.97619</td>
<td>0.25</td>
</tr>
<tr>
<td>rs9310888</td>
<td>3</td>
<td>50.85161</td>
<td>0.075</td>
<td>0</td>
</tr>
<tr>
<td>rs10510791</td>
<td>3</td>
<td>79.71704</td>
<td>0.487805</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs1395771</td>
<td>3</td>
<td>120.2731</td>
<td>0.02381</td>
<td>0.566667</td>
</tr>
<tr>
<td>rs1919550</td>
<td>3</td>
<td>141.8274</td>
<td>0.02381</td>
<td>0.833333</td>
</tr>
<tr>
<td>rs2035573</td>
<td>3</td>
<td>152.983</td>
<td>0.654762</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs1984473</td>
<td>3</td>
<td>179.4574</td>
<td>0.630952</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs6804094</td>
<td>3</td>
<td>221.6831</td>
<td>0.653846</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs1398829</td>
<td>4</td>
<td>40.00216</td>
<td>0.97619</td>
<td>1</td>
</tr>
<tr>
<td>rs719776</td>
<td>4</td>
<td>55.62262</td>
<td>0.142857</td>
<td>0.166667</td>
</tr>
<tr>
<td>rs10517518</td>
<td>4</td>
<td>79.25042</td>
<td>0.857143</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs9307613</td>
<td>4</td>
<td>137.4881</td>
<td>0.775</td>
<td>0.833333</td>
</tr>
<tr>
<td>rs10519979</td>
<td>4</td>
<td>163.4903</td>
<td>0.45122</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs10520440</td>
<td>4</td>
<td>204.5646</td>
<td>0.785714</td>
<td>0.166667</td>
</tr>
<tr>
<td>rs</td>
<td>n</td>
<td>s</td>
<td>g1</td>
<td>g2</td>
</tr>
<tr>
<td>----------</td>
<td>---</td>
<td>---</td>
<td>-------</td>
<td>----</td>
</tr>
<tr>
<td>rs257748</td>
<td>5</td>
<td>34.79072</td>
<td>0.380952</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs1353251</td>
<td>5</td>
<td>57.95357</td>
<td>0.797619</td>
<td>0.266667</td>
</tr>
<tr>
<td>rs9292118</td>
<td>5</td>
<td>70.44465</td>
<td>0.833333</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs6883095</td>
<td>5</td>
<td>94.2854</td>
<td>0.547619</td>
<td>0.033333</td>
</tr>
<tr>
<td>rs153898</td>
<td>5</td>
<td>106.4104</td>
<td>0.738095</td>
<td>0.033333</td>
</tr>
<tr>
<td>rs1990745</td>
<td>5</td>
<td>113.4173</td>
<td>0.880952</td>
<td>0.233333</td>
</tr>
<tr>
<td>rs10515535</td>
<td>5</td>
<td>157.4409</td>
<td>0.285714</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs2569029</td>
<td>5</td>
<td>172.031</td>
<td>0.666667</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs1477277</td>
<td>5</td>
<td>217.2694</td>
<td>0.345238</td>
<td>0.3</td>
</tr>
<tr>
<td>rs6911727</td>
<td>6</td>
<td>20.75571</td>
<td>0.47619</td>
<td>1</td>
</tr>
<tr>
<td>rs10484578</td>
<td>6</td>
<td>54.20572</td>
<td>0.375</td>
<td>0.066667</td>
</tr>
<tr>
<td>rs993314</td>
<td>6</td>
<td>95.56578</td>
<td>0.166667</td>
<td>0.7</td>
</tr>
<tr>
<td>rs10498919</td>
<td>6</td>
<td>98.31055</td>
<td>0.833333</td>
<td>0.2</td>
</tr>
<tr>
<td>rs10491654</td>
<td>9</td>
<td>123.4304</td>
<td>0.52381</td>
<td>0.9</td>
</tr>
<tr>
<td>rs2585901</td>
<td>12</td>
<td>86.83803</td>
<td>0.973684</td>
<td>1</td>
</tr>
<tr>
<td>rs10506816</td>
<td>12</td>
<td>96.8146</td>
<td>0.642857</td>
<td>0.8</td>
</tr>
<tr>
<td>rs249847</td>
<td>12</td>
<td>118.7013</td>
<td>0.642857</td>
<td>0.8</td>
</tr>
<tr>
<td>rs407600</td>
<td>12</td>
<td>141.342</td>
<td>0.833333</td>
<td>0.9</td>
</tr>
<tr>
<td>rs4034627</td>
<td>12</td>
<td>165.957</td>
<td>0.047619</td>
<td>0.0</td>
</tr>
<tr>
<td>rs2585901</td>
<td>13</td>
<td>167.0217</td>
<td>0.654762</td>
<td>1</td>
</tr>
<tr>
<td>rs6569792</td>
<td>6</td>
<td>148.7193</td>
<td>0.678571</td>
<td>0.8</td>
</tr>
<tr>
<td>rs257748</td>
<td>5</td>
<td>34.79072</td>
<td>0.380952</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs1353251</td>
<td>5</td>
<td>57.95357</td>
<td>0.797619</td>
<td>0.266667</td>
</tr>
<tr>
<td>rs9292118</td>
<td>5</td>
<td>70.44465</td>
<td>0.833333</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs6883095</td>
<td>5</td>
<td>94.2854</td>
<td>0.547619</td>
<td>0.033333</td>
</tr>
<tr>
<td>rs153898</td>
<td>5</td>
<td>106.4104</td>
<td>0.738095</td>
<td>0.033333</td>
</tr>
<tr>
<td>rs1990745</td>
<td>5</td>
<td>113.4173</td>
<td>0.880952</td>
<td>0.233333</td>
</tr>
<tr>
<td>rs10515535</td>
<td>5</td>
<td>157.4409</td>
<td>0.285714</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs2569029</td>
<td>5</td>
<td>172.031</td>
<td>0.666667</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs1477277</td>
<td>5</td>
<td>217.2694</td>
<td>0.345238</td>
<td>0.3</td>
</tr>
<tr>
<td>rs6911727</td>
<td>6</td>
<td>20.75571</td>
<td>0.47619</td>
<td>1</td>
</tr>
<tr>
<td>rs10484578</td>
<td>6</td>
<td>54.20572</td>
<td>0.375</td>
<td>0.066667</td>
</tr>
<tr>
<td>rs993314</td>
<td>6</td>
<td>95.56578</td>
<td>0.166667</td>
<td>0.7</td>
</tr>
<tr>
<td>rs10498919</td>
<td>6</td>
<td>98.31055</td>
<td>0.833333</td>
<td>0.2</td>
</tr>
<tr>
<td>rs9320808</td>
<td>6</td>
<td>137.3366</td>
<td>0.095238</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs6569792</td>
<td>6</td>
<td>148.7193</td>
<td>0.678571</td>
<td>0.8</td>
</tr>
<tr>
<td>rs9295817</td>
<td>7</td>
<td>150.3822</td>
<td>0.785714</td>
<td>0.2</td>
</tr>
<tr>
<td>rs10248051</td>
<td>7</td>
<td>167.0217</td>
<td>0.654762</td>
<td>1</td>
</tr>
<tr>
<td>rs10214949</td>
<td>7</td>
<td>191.8366</td>
<td>0.833333</td>
<td>0.2</td>
</tr>
<tr>
<td>rs10488172</td>
<td>7</td>
<td>20.75571</td>
<td>0.47619</td>
<td>1</td>
</tr>
<tr>
<td>rs802524</td>
<td>7</td>
<td>54.20572</td>
<td>0.375</td>
<td>0.066667</td>
</tr>
<tr>
<td>rs9325872</td>
<td>8</td>
<td>95.56578</td>
<td>0.166667</td>
<td>0.7</td>
</tr>
<tr>
<td>rs7463344</td>
<td>8</td>
<td>98.31055</td>
<td>0.833333</td>
<td>0.2</td>
</tr>
<tr>
<td>rs4130405</td>
<td>8</td>
<td>122.0503</td>
<td>0.190476</td>
<td>0.8</td>
</tr>
<tr>
<td>rs1898280</td>
<td>8</td>
<td>134.5748</td>
<td>0.238095</td>
<td>0.833333</td>
</tr>
<tr>
<td>rs4733652</td>
<td>8</td>
<td>154.9088</td>
<td>0.761905</td>
<td>0.1</td>
</tr>
<tr>
<td>rs2840290</td>
<td>9</td>
<td>172.031</td>
<td>0.666667</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs879780</td>
<td>11</td>
<td>126.861</td>
<td>0.845238</td>
<td>0.666667</td>
</tr>
<tr>
<td>rs10506816</td>
<td>12</td>
<td>96.8146</td>
<td>0.642857</td>
<td>0.8</td>
</tr>
<tr>
<td>rs249847</td>
<td>12</td>
<td>118.7013</td>
<td>0.463415</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs407600</td>
<td>12</td>
<td>141.342</td>
<td>0.833333</td>
<td>0.9</td>
</tr>
<tr>
<td>rs4034627</td>
<td>12</td>
<td>165.957</td>
<td>0.047619</td>
<td>0.0</td>
</tr>
<tr>
<td>rs2585901</td>
<td>13</td>
<td>167.0217</td>
<td>0.654762</td>
<td>1</td>
</tr>
<tr>
<td>rs10501474</td>
<td>11</td>
<td>106.6125</td>
<td>0.966667</td>
<td>0.9</td>
</tr>
<tr>
<td>rs5000507</td>
<td>13</td>
<td>86.49679</td>
<td>0.7125</td>
<td>1</td>
</tr>
<tr>
<td>rs10492585</td>
<td>13</td>
<td>112.5483</td>
<td>0.940476</td>
<td>1</td>
</tr>
<tr>
<td>rs9323178</td>
<td>14</td>
<td>9.523717</td>
<td>0.452381</td>
<td>0.966667</td>
</tr>
<tr>
<td>Marker ID</td>
<td>Chr</td>
<td>cM</td>
<td>European Frequency</td>
<td>Native American Frequency</td>
</tr>
<tr>
<td>-----------</td>
<td>-----</td>
<td>-------</td>
<td>--------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>rs1451928</td>
<td>14</td>
<td>44.07544</td>
<td>0.857143</td>
<td>0.2</td>
</tr>
<tr>
<td>rs2296274</td>
<td>14</td>
<td>58.42676</td>
<td>0.75</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs10131076</td>
<td>14</td>
<td>78.56427</td>
<td>0.071429</td>
<td>0</td>
</tr>
<tr>
<td>rs9302185</td>
<td>15</td>
<td>48.97097</td>
<td>0.190476</td>
<td>0.033333</td>
</tr>
<tr>
<td>rs10520678</td>
<td>15</td>
<td>95.60323</td>
<td>0.731707</td>
<td>1</td>
</tr>
<tr>
<td>rs30125</td>
<td>16</td>
<td>38.20662</td>
<td>0.04878</td>
<td>0.1</td>
</tr>
<tr>
<td>rs1004704</td>
<td>16</td>
<td>62.95409</td>
<td>0.214286</td>
<td>0.866667</td>
</tr>
<tr>
<td>rs10500505</td>
<td>16</td>
<td>87.59216</td>
<td>0.214286</td>
<td>0.8</td>
</tr>
<tr>
<td>rs30125</td>
<td>16</td>
<td>106.413</td>
<td>0.926829</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs9302185</td>
<td>15</td>
<td>99.60422</td>
<td>0.083333</td>
<td>0.192308</td>
</tr>
<tr>
<td>rs10491097</td>
<td>17</td>
<td>47.36648</td>
<td>0.647059</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs225624</td>
<td>17</td>
<td>123.8847</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>rs1013459</td>
<td>18</td>
<td>38.94166</td>
<td>0.9</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs2042762</td>
<td>18</td>
<td>64.41584</td>
<td>0</td>
<td>0.733333</td>
</tr>
<tr>
<td>rs12953952</td>
<td>18</td>
<td>106.413</td>
<td>0.926829</td>
<td>0.966667</td>
</tr>
<tr>
<td>rs888861</td>
<td>19</td>
<td>70.13325</td>
<td>0.738095</td>
<td>0.9</td>
</tr>
<tr>
<td>rs798887</td>
<td>19</td>
<td>104.1811</td>
<td>0.853659</td>
<td>0.133333</td>
</tr>
<tr>
<td>rs708915</td>
<td>20</td>
<td>23.80641</td>
<td>0.1</td>
<td>0.733333</td>
</tr>
<tr>
<td>rs10500505</td>
<td>16</td>
<td>106.5914</td>
<td>0.642857</td>
<td>0.766667</td>
</tr>
<tr>
<td>rs2829454</td>
<td>21</td>
<td>23.73772</td>
<td>0.27381</td>
<td>0.933333</td>
</tr>
<tr>
<td>rs130222</td>
<td>22</td>
<td>53.71687</td>
<td>0.261905</td>
<td>0.033333</td>
</tr>
</tbody>
</table>

**Bonilla AIMS Panel (Used for San Luis Valley and Guerrero Samples)**
<table>
<thead>
<tr>
<th>Marker ID</th>
<th>Chr</th>
<th>cM</th>
<th>European Frequency</th>
<th>Native American Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs263531</td>
<td>1</td>
<td>5.9</td>
<td>0.543</td>
<td>0.086</td>
</tr>
<tr>
<td>rs434504</td>
<td>1</td>
<td>14.73</td>
<td>0.871</td>
<td>0.681</td>
</tr>
<tr>
<td>rs1039630</td>
<td>1</td>
<td>62.69</td>
<td>0.298</td>
<td>0.576</td>
</tr>
<tr>
<td>rs950848</td>
<td>1</td>
<td>66.84</td>
<td>0.017</td>
<td>0.19</td>
</tr>
<tr>
<td>rs595961</td>
<td>1</td>
<td>66.91</td>
<td>0.833</td>
<td>0.441</td>
</tr>
<tr>
<td>rs522287</td>
<td>1</td>
<td>83.13</td>
<td>0.574</td>
<td>0.061</td>
</tr>
<tr>
<td>rs1469344</td>
<td>1</td>
<td>86.48</td>
<td>0.613</td>
<td>0.981</td>
</tr>
<tr>
<td>rs625994</td>
<td>1</td>
<td>91.64</td>
<td>0.912</td>
<td>0.39</td>
</tr>
<tr>
<td>rs1325609</td>
<td>1</td>
<td>101.29</td>
<td>0.647</td>
<td>0.754</td>
</tr>
<tr>
<td>rs783064</td>
<td>1</td>
<td>102.47</td>
<td>0.466</td>
<td>0.483</td>
</tr>
<tr>
<td>rs1937025</td>
<td>1</td>
<td>112.2</td>
<td>0.035</td>
<td>0.22</td>
</tr>
<tr>
<td>rs236336</td>
<td>1</td>
<td>126.99</td>
<td>0.718</td>
<td>0.991</td>
</tr>
<tr>
<td>rs841338</td>
<td>1</td>
<td>128.64</td>
<td>0.966</td>
<td>0.828</td>
</tr>
<tr>
<td>rs2814778</td>
<td>1</td>
<td>164.82</td>
<td>0.009</td>
<td>0.008</td>
</tr>
<tr>
<td>rs1446966</td>
<td>1</td>
<td>164.99</td>
<td>0.429</td>
<td>0.422</td>
</tr>
<tr>
<td>rs1415680</td>
<td>1</td>
<td>181.4</td>
<td>0.327</td>
<td>0.136</td>
</tr>
<tr>
<td>rs6003</td>
<td>1</td>
<td>211.97</td>
<td>0.93</td>
<td>0.924</td>
</tr>
<tr>
<td>rs2204307</td>
<td>1</td>
<td>214.64</td>
<td>0.143</td>
<td>0.22</td>
</tr>
<tr>
<td>rs296528</td>
<td>1</td>
<td>214.7</td>
<td>0.333</td>
<td>0.017</td>
</tr>
<tr>
<td>rs2065160</td>
<td>1</td>
<td>222.38</td>
<td>0.053</td>
<td>0.922</td>
</tr>
<tr>
<td>rs1406869</td>
<td>1</td>
<td>251.21</td>
<td>0.353</td>
<td>0.703</td>
</tr>
<tr>
<td>rs715956</td>
<td>1</td>
<td>268.05</td>
<td>0.214</td>
<td>0.339</td>
</tr>
<tr>
<td>rs772436</td>
<td>2</td>
<td>17.22</td>
<td>0.281</td>
<td>0.737</td>
</tr>
<tr>
<td>rs1869380</td>
<td>2</td>
<td>29.14</td>
<td>0.362</td>
<td>0.636</td>
</tr>
<tr>
<td>rs997676</td>
<td>2</td>
<td>40.6</td>
<td>0.109</td>
<td>0.127</td>
</tr>
<tr>
<td>rs892457</td>
<td>2</td>
<td>58.61</td>
<td>0.473</td>
<td>0.603</td>
</tr>
<tr>
<td>rs1431332</td>
<td>2</td>
<td>63.37</td>
<td>0.518</td>
<td>0.983</td>
</tr>
<tr>
<td>rs3287</td>
<td>2</td>
<td>79.31</td>
<td>0.723</td>
<td>0.864</td>
</tr>
<tr>
<td>rs727878</td>
<td>2</td>
<td>80.42</td>
<td>0.255</td>
<td>0.28</td>
</tr>
<tr>
<td>rs262482</td>
<td>2</td>
<td>85.52</td>
<td>0.324</td>
<td>0.009</td>
</tr>
<tr>
<td>rs53915</td>
<td>2</td>
<td>105.87</td>
<td>0.655</td>
<td>0.305</td>
</tr>
<tr>
<td>rs1221172</td>
<td>2</td>
<td>107.42</td>
<td>0.553</td>
<td>0.136</td>
</tr>
<tr>
<td>rs959929</td>
<td>2</td>
<td>113.57</td>
<td>0.491</td>
<td>0.119</td>
</tr>
<tr>
<td>rs</td>
<td>pos</td>
<td>p</td>
<td>MAF</td>
<td>p additive</td>
</tr>
<tr>
<td>--------------</td>
<td>-----</td>
<td>----</td>
<td>-----</td>
<td>------------</td>
</tr>
<tr>
<td>rs729253</td>
<td>2</td>
<td>132.44</td>
<td>0.6</td>
<td>0.966</td>
</tr>
<tr>
<td>rs725395</td>
<td>2</td>
<td>138.48</td>
<td>0.426</td>
<td>0.84</td>
</tr>
<tr>
<td>rs1447111</td>
<td>2</td>
<td>228.63</td>
<td>0.681</td>
<td>0.881</td>
</tr>
<tr>
<td>rs1848728</td>
<td>2</td>
<td>240.71</td>
<td>0.625</td>
<td>0.921</td>
</tr>
<tr>
<td>rs1533677</td>
<td>3</td>
<td>67.6</td>
<td>0.472</td>
<td>0.732</td>
</tr>
<tr>
<td>rs1108718</td>
<td>3</td>
<td>92.39</td>
<td>0.579</td>
<td>0.822</td>
</tr>
<tr>
<td>rs830599</td>
<td>3</td>
<td>99.37</td>
<td>0.31</td>
<td>0.966</td>
</tr>
<tr>
<td>rs1399272</td>
<td>3</td>
<td>122.59</td>
<td>0.595</td>
<td>0.229</td>
</tr>
<tr>
<td>rs1105220</td>
<td>3</td>
<td>141.64</td>
<td>0</td>
<td>0.009</td>
</tr>
<tr>
<td>rs361065</td>
<td>3</td>
<td>159.99</td>
<td>0.405</td>
<td>0.017</td>
</tr>
<tr>
<td>rs361055</td>
<td>3</td>
<td>160</td>
<td>0.404</td>
<td>0.017</td>
</tr>
<tr>
<td>rs892263</td>
<td>3</td>
<td>167.09</td>
<td>0.11</td>
<td>0.069</td>
</tr>
<tr>
<td>rs715790</td>
<td>3</td>
<td>224.11</td>
<td>0.42</td>
<td>0.575</td>
</tr>
<tr>
<td>rs2045517</td>
<td>4</td>
<td>103.22</td>
<td>0.675</td>
<td>0.526</td>
</tr>
<tr>
<td>rs959858</td>
<td>4</td>
<td>118.51</td>
<td>0.482</td>
<td>0.22</td>
</tr>
<tr>
<td>rs733563</td>
<td>4</td>
<td>170.67</td>
<td>0.661</td>
<td>0.627</td>
</tr>
<tr>
<td>rs997164</td>
<td>4</td>
<td>200.49</td>
<td>0.58</td>
<td>1</td>
</tr>
<tr>
<td>rs1368872</td>
<td>4</td>
<td>202.22</td>
<td>0.517</td>
<td>0.14</td>
</tr>
<tr>
<td>rs1501680</td>
<td>5</td>
<td>48.51</td>
<td>0.774</td>
<td>0.454</td>
</tr>
<tr>
<td>rs270565</td>
<td>5</td>
<td>60.48</td>
<td>0.912</td>
<td>1</td>
</tr>
<tr>
<td>rs1125508</td>
<td>5</td>
<td>63.13</td>
<td>0.31</td>
<td>0.25</td>
</tr>
<tr>
<td>rs875543</td>
<td>5</td>
<td>87.89</td>
<td>0.842</td>
<td>0.693</td>
</tr>
<tr>
<td>rs3317</td>
<td>5</td>
<td>128.75</td>
<td>0.5</td>
<td>0.653</td>
</tr>
<tr>
<td>rs1005056</td>
<td>5</td>
<td>129.83</td>
<td>0.064</td>
<td>0.455</td>
</tr>
<tr>
<td>rs960709</td>
<td>5</td>
<td>168.99</td>
<td>0.717</td>
<td>0.375</td>
</tr>
<tr>
<td>rs3340</td>
<td>5</td>
<td>171.28</td>
<td>0.211</td>
<td>0.788</td>
</tr>
<tr>
<td>rs925197</td>
<td>5</td>
<td>207.24</td>
<td>0.063</td>
<td>0.172</td>
</tr>
<tr>
<td>rs1407361</td>
<td>6</td>
<td>18.9</td>
<td>0.318</td>
<td>0.644</td>
</tr>
<tr>
<td>rs933199</td>
<td>6</td>
<td>48.98</td>
<td>0.902</td>
<td>0.729</td>
</tr>
<tr>
<td>rs559035</td>
<td>6</td>
<td>78.74</td>
<td>0.853</td>
<td>0.552</td>
</tr>
<tr>
<td>rs1937147</td>
<td>6</td>
<td>83.96</td>
<td>0.526</td>
<td>0.941</td>
</tr>
<tr>
<td>rs718268</td>
<td>6</td>
<td>121.37</td>
<td>0.684</td>
<td>0.432</td>
</tr>
<tr>
<td>rs1041656</td>
<td>6</td>
<td>162.7</td>
<td>0.622</td>
<td>0.313</td>
</tr>
<tr>
<td>rs765338</td>
<td>6</td>
<td>175.11</td>
<td>0.898</td>
<td>0.791</td>
</tr>
<tr>
<td>rs877823</td>
<td>6</td>
<td>192.18</td>
<td>0.737</td>
<td>0.958</td>
</tr>
<tr>
<td>rs409359</td>
<td>6</td>
<td>194.3</td>
<td>0.529</td>
<td>0.51</td>
</tr>
<tr>
<td>rs1476597</td>
<td>7</td>
<td>49.4</td>
<td>0.411</td>
<td>0.449</td>
</tr>
<tr>
<td>rs1155513</td>
<td>7</td>
<td>60.52</td>
<td>0.526</td>
<td>0.982</td>
</tr>
<tr>
<td>rs984654</td>
<td>7</td>
<td>77.69</td>
<td>0.167</td>
<td>0.695</td>
</tr>
<tr>
<td>rs1018919</td>
<td>7</td>
<td>101.47</td>
<td>0.25</td>
<td>0.293</td>
</tr>
<tr>
<td>rs2161</td>
<td>7</td>
<td>110.08</td>
<td>0.295</td>
<td>0.595</td>
</tr>
<tr>
<td>rs1528037</td>
<td>7</td>
<td>131.46</td>
<td>0.596</td>
<td>0.534</td>
</tr>
<tr>
<td>rs125097</td>
<td>7</td>
<td>148.63</td>
<td>0.453</td>
<td>0.281</td>
</tr>
<tr>
<td>rs869337</td>
<td>7</td>
<td>178.8</td>
<td>0.457</td>
<td>0</td>
</tr>
<tr>
<td>rs716840</td>
<td>8</td>
<td>7.25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rs1454284</td>
<td>8</td>
<td>13.99</td>
<td>0.409</td>
<td>0.974</td>
</tr>
<tr>
<td>rs1039917</td>
<td>8</td>
<td>20.97</td>
<td>0.375</td>
<td>0.513</td>
</tr>
<tr>
<td>rs285</td>
<td>8</td>
<td>48.7</td>
<td>0.569</td>
<td>0.629</td>
</tr>
<tr>
<td>rs351782</td>
<td>8</td>
<td>62.17</td>
<td>0.94</td>
<td>0.959</td>
</tr>
<tr>
<td>rs</td>
<td>Chromosome</td>
<td>Value</td>
<td>Probability</td>
<td>p-value</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
<td>-------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>rs3176921</td>
<td>8</td>
<td>93.94</td>
<td>0.079</td>
<td>0.00</td>
</tr>
<tr>
<td>rs717836</td>
<td>8</td>
<td>134.2</td>
<td>0.132</td>
<td>0.001</td>
</tr>
<tr>
<td>rs1467044</td>
<td>8</td>
<td>138.46</td>
<td>0.588</td>
<td>0.802</td>
</tr>
<tr>
<td>rs913258</td>
<td>9</td>
<td>13.07</td>
<td>0.862</td>
<td>0.754</td>
</tr>
<tr>
<td>rs488952</td>
<td>9</td>
<td>33.43</td>
<td>0.56</td>
<td>0.805</td>
</tr>
<tr>
<td>rs998599</td>
<td>9</td>
<td>35.7</td>
<td>0.43</td>
<td>0.949</td>
</tr>
<tr>
<td>rs1414241</td>
<td>9</td>
<td>42.07</td>
<td>0.552</td>
<td>0.839</td>
</tr>
<tr>
<td>rs721702</td>
<td>9</td>
<td>43.45</td>
<td>0.825</td>
<td>0.839</td>
</tr>
<tr>
<td>rs1041321</td>
<td>9</td>
<td>55.83</td>
<td>0.377</td>
<td>0.167</td>
</tr>
<tr>
<td>rs2695</td>
<td>9</td>
<td>87.02</td>
<td>0.267</td>
<td>0.907</td>
</tr>
<tr>
<td>rs526454</td>
<td>9</td>
<td>98.87</td>
<td>0.944</td>
<td>0.819</td>
</tr>
<tr>
<td>rs675837</td>
<td>9</td>
<td>99.07</td>
<td>0.33</td>
<td>0.61</td>
</tr>
<tr>
<td>rs662117</td>
<td>9</td>
<td>161.1</td>
<td>0.911</td>
<td>0.422</td>
</tr>
<tr>
<td>rs9032</td>
<td>9</td>
<td>163.16</td>
<td>0.904</td>
<td>0.415</td>
</tr>
<tr>
<td>rs590086</td>
<td>9</td>
<td>174.6</td>
<td>0.888</td>
<td>1.00</td>
</tr>
<tr>
<td>rs1076160</td>
<td>9</td>
<td>179.57</td>
<td>0.518</td>
<td>0.377</td>
</tr>
<tr>
<td>rs723220</td>
<td>10</td>
<td>13.66</td>
<td>0.667</td>
<td>0.093</td>
</tr>
<tr>
<td>rs913375</td>
<td>10</td>
<td>28.78</td>
<td>0.421</td>
<td>0.974</td>
</tr>
<tr>
<td>rs721825</td>
<td>10</td>
<td>73.02</td>
<td>0.241</td>
<td>0.398</td>
</tr>
<tr>
<td>rs1983128</td>
<td>10</td>
<td>88.81</td>
<td>0.645</td>
<td>0.862</td>
</tr>
<tr>
<td>rs877783</td>
<td>10</td>
<td>95.22</td>
<td>0.816</td>
<td>0.56</td>
</tr>
<tr>
<td>rs1034290</td>
<td>10</td>
<td>96.31</td>
<td>0.652</td>
<td>0.293</td>
</tr>
<tr>
<td>rs719909</td>
<td>10</td>
<td>119.66</td>
<td>0.857</td>
<td>0.776</td>
</tr>
<tr>
<td>rs1546541</td>
<td>11</td>
<td>21.48</td>
<td>0.202</td>
<td>0.576</td>
</tr>
<tr>
<td>rs1385851</td>
<td>11</td>
<td>25.95</td>
<td>0.842</td>
<td>1.00</td>
</tr>
<tr>
<td>rs1470144</td>
<td>11</td>
<td>30.04</td>
<td>0.667</td>
<td>0.195</td>
</tr>
<tr>
<td>rs713503</td>
<td>11</td>
<td>31.17</td>
<td>0.509</td>
<td>0.873</td>
</tr>
<tr>
<td>rs523200</td>
<td>11</td>
<td>85.89</td>
<td>0.107</td>
<td>0.457</td>
</tr>
<tr>
<td>rs594689</td>
<td>11</td>
<td>86.53</td>
<td>0.578</td>
<td>0.153</td>
</tr>
<tr>
<td>rs725192</td>
<td>11</td>
<td>108.57</td>
<td>0.17</td>
<td>0.00</td>
</tr>
<tr>
<td>rs282496</td>
<td>11</td>
<td>122.27</td>
<td>0.728</td>
<td>0.958</td>
</tr>
<tr>
<td>rs1800498</td>
<td>11</td>
<td>136.87</td>
<td>0.325</td>
<td>0.992</td>
</tr>
<tr>
<td>rs236919</td>
<td>11</td>
<td>144.05</td>
<td>0.621</td>
<td>0.237</td>
</tr>
<tr>
<td>rs901892</td>
<td>11</td>
<td>144.05</td>
<td>0.848</td>
<td>0.602</td>
</tr>
<tr>
<td>rs672122</td>
<td>12</td>
<td>124.04</td>
<td>0.5</td>
<td>0.216</td>
</tr>
<tr>
<td>rs594226</td>
<td>12</td>
<td>133.52</td>
<td>0.764</td>
<td>0.402</td>
</tr>
<tr>
<td>rs1407961</td>
<td>13</td>
<td>0</td>
<td>0.914</td>
<td>0.965</td>
</tr>
<tr>
<td>rs883055</td>
<td>13</td>
<td>21.55</td>
<td>0.054</td>
<td>0.203</td>
</tr>
<tr>
<td>rs1337038</td>
<td>13</td>
<td>79.16</td>
<td>0.518</td>
<td>0.144</td>
</tr>
<tr>
<td>rs987284</td>
<td>13</td>
<td>84.73</td>
<td>0.56</td>
<td>0.085</td>
</tr>
<tr>
<td>rs320075</td>
<td>13</td>
<td>94.04</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>rs1113337</td>
<td>13</td>
<td>132.05</td>
<td>0.595</td>
<td>0.421</td>
</tr>
<tr>
<td>rs1998055</td>
<td>14</td>
<td>10.11</td>
<td>0.754</td>
<td>1.00</td>
</tr>
<tr>
<td>rs729531</td>
<td>14</td>
<td>10.37</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rs953743</td>
<td>14</td>
<td>49.74</td>
<td>0.519</td>
<td>0.153</td>
</tr>
<tr>
<td>rs1111108</td>
<td>14</td>
<td>58.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>rs915056</td>
<td>14</td>
<td>62.67</td>
<td>0.664</td>
<td>0.362</td>
</tr>
<tr>
<td>rs763807</td>
<td>14</td>
<td>64.21</td>
<td>0.145</td>
<td>0.289</td>
</tr>
<tr>
<td>rs1007407</td>
<td>14</td>
<td>64.29</td>
<td>0.675</td>
<td>0.008</td>
</tr>
<tr>
<td>rs1157223</td>
<td>14</td>
<td>81.59</td>
<td>0.107</td>
<td>0</td>
</tr>
<tr>
<td>rs174518</td>
<td>14</td>
<td>84.68</td>
<td>0.482</td>
<td>0.353</td>
</tr>
<tr>
<td>rs1296149</td>
<td>14</td>
<td>84.69</td>
<td>0.5</td>
<td>0.397</td>
</tr>
<tr>
<td>rs741272</td>
<td>14</td>
<td>87.65</td>
<td>0.509</td>
<td>0.186</td>
</tr>
<tr>
<td>rs716873</td>
<td>14</td>
<td>92.35</td>
<td>0.193</td>
<td>0.017</td>
</tr>
<tr>
<td>rs1076808</td>
<td>14</td>
<td>95.62</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>rs974324</td>
<td>14</td>
<td>113.49</td>
<td>1</td>
<td>0.983</td>
</tr>
<tr>
<td>rs730570</td>
<td>14</td>
<td>118.49</td>
<td>0.897</td>
<td>0.068</td>
</tr>
<tr>
<td>rs1426217</td>
<td>15</td>
<td>9.68</td>
<td>0.366</td>
<td>0.593</td>
</tr>
<tr>
<td>rs1426208</td>
<td>15</td>
<td>9.9</td>
<td>0.333</td>
<td>0.543</td>
</tr>
<tr>
<td>rs2862</td>
<td>15</td>
<td>29.5</td>
<td>0.216</td>
<td>0.712</td>
</tr>
<tr>
<td>rs920915</td>
<td>15</td>
<td>52.98</td>
<td>0.517</td>
<td>0.678</td>
</tr>
<tr>
<td>rs972801</td>
<td>15</td>
<td>54.37</td>
<td>0.219</td>
<td>0.763</td>
</tr>
<tr>
<td>rs2010069</td>
<td>15</td>
<td>89.46</td>
<td>0.412</td>
<td>0.839</td>
</tr>
<tr>
<td>rs1015081</td>
<td>15</td>
<td>96.65</td>
<td>0.431</td>
<td>0.441</td>
</tr>
<tr>
<td>rs1375229</td>
<td>15</td>
<td>121.33</td>
<td>0.021</td>
<td>0</td>
</tr>
<tr>
<td>rs251741</td>
<td>16</td>
<td>13.41</td>
<td>0.241</td>
<td>0.149</td>
</tr>
<tr>
<td>rs1395580</td>
<td>16</td>
<td>21.62</td>
<td>0.825</td>
<td>0.864</td>
</tr>
<tr>
<td>rs1395579</td>
<td>16</td>
<td>21.62</td>
<td>0.66</td>
<td>0.06</td>
</tr>
<tr>
<td>rs1650999</td>
<td>16</td>
<td>34.3</td>
<td>0.964</td>
<td>0.364</td>
</tr>
<tr>
<td>rs878671</td>
<td>16</td>
<td>56.14</td>
<td>0.377</td>
<td>0.589</td>
</tr>
<tr>
<td>rs173537</td>
<td>16</td>
<td>77.92</td>
<td>0.246</td>
<td>0.331</td>
</tr>
<tr>
<td>rs67302</td>
<td>16</td>
<td>88.14</td>
<td>0.422</td>
<td>0.771</td>
</tr>
<tr>
<td>rs886895</td>
<td>17</td>
<td>29.81341</td>
<td>0.144</td>
<td>0.5</td>
</tr>
<tr>
<td>rs212498</td>
<td>17</td>
<td>63.21</td>
<td>0.772</td>
<td>0.629</td>
</tr>
<tr>
<td>rs917502</td>
<td>17</td>
<td>63.57</td>
<td>0.483</td>
<td>0.982</td>
</tr>
<tr>
<td>rs712047</td>
<td>17</td>
<td>65.61945</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>rs712037</td>
<td>17</td>
<td>68.00131</td>
<td>0.85</td>
<td>0.5</td>
</tr>
<tr>
<td>rs758767</td>
<td>17</td>
<td>114.24</td>
<td>0.833</td>
<td>0.585</td>
</tr>
<tr>
<td>rs1030525</td>
<td>18</td>
<td>50.09</td>
<td>0.009</td>
<td>0.457</td>
</tr>
<tr>
<td>rs1125425</td>
<td>18</td>
<td>62.79</td>
<td>0.104</td>
<td>0.357</td>
</tr>
<tr>
<td>rs1437069</td>
<td>18</td>
<td>82.93</td>
<td>0.228</td>
<td>0.551</td>
</tr>
<tr>
<td>rs667508</td>
<td>18</td>
<td>83.26</td>
<td>0.821</td>
<td>0.797</td>
</tr>
<tr>
<td>rs1432065</td>
<td>18</td>
<td>109.88</td>
<td>0.777</td>
<td>0.912</td>
</tr>
<tr>
<td>rs1465708</td>
<td>19</td>
<td>46.41</td>
<td>0.315</td>
<td>0.712</td>
</tr>
<tr>
<td>rs886993</td>
<td>19</td>
<td>75.69936</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>rs430667</td>
<td>19</td>
<td>79.69</td>
<td>0.57</td>
<td>0.085</td>
</tr>
<tr>
<td>rs275837</td>
<td>19</td>
<td>84.52</td>
<td>0.483</td>
<td>0.186</td>
</tr>
<tr>
<td>rs1040577</td>
<td>20</td>
<td>47.2</td>
<td>0.465</td>
<td>0.22</td>
</tr>
<tr>
<td>rs1004571</td>
<td>20</td>
<td>73.96</td>
<td>0.345</td>
<td>0.431</td>
</tr>
<tr>
<td>rs172982</td>
<td>20</td>
<td>96.19</td>
<td>0.454</td>
<td>0.5</td>
</tr>
<tr>
<td>rs735050</td>
<td>20</td>
<td>96.2</td>
<td>0.784</td>
<td>0.72</td>
</tr>
<tr>
<td>rs2014519</td>
<td>21</td>
<td>16.76</td>
<td>0.296</td>
<td>0.39</td>
</tr>
<tr>
<td>rs81481</td>
<td>21</td>
<td>53.88</td>
<td>0.175</td>
<td>0.06</td>
</tr>
<tr>
<td>rs138335</td>
<td>22</td>
<td>54.02</td>
<td>0.333</td>
<td>0.008</td>
</tr>
</tbody>
</table>
VITA
Ellen E. Quillen

Education

Pennsylvania State University, PhD, Anthropology, December 2010
Dissertation Title: Investigating genes related to the evolution of Indigenous American skin pigmentation.

Pennsylvania State University, MA, with distinction, Anthropology, May 2007

University of Kansas, BS, with honors, Genetics, May 2005
Thesis Title: Genetic differentiation in Newfoundland outports.

Professional Memberships
American Society of Human Genetics, American Association of Physical Anthropology, American Association of Anthropological Genetics

Grants and Awards

2010 Aleš Hrdlička Award for Outstanding Student Paper at AAPA Meeting
2010 American Association of Anthropological Genetics Outstanding Student Presentation Prize
2009 NSF Doctoral Dissertation Improvement Grant # 0925976
2009 College of Liberal Arts RGSO Dissertation Support Grant, Pennsylvania State University

Publications


Select Abstracts


Teaching Experience and Service

Pennsylvania State University, Department of Anthropology. Instructor for introductory course in laboratory methods and introductory biological anthropology course.
Mentor Undergraduate and Master’s Researchers, Anthropological Genomics Lab, Pennsylvania State University, 2005-present
Developed workshop for Penn State Women in Science and Engineering Summer Camp for high school girls, June 2009, November 2010
President, Anthropology Graduate Student Association, Pennsylvania State University, 2007