MEASURING AND USING INDIVIDUAL GENOMIC ANCESTRY
TO STUDY COMPLEX PHENOTYPES

A Thesis in
Genetics
by
Indrani Halder

© 2005 Indrani Halder

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2005
The thesis of Indrani Halder was reviewed and approved* by the following:

Mark D Shriver  
Associate Professor of Biological Anthropology and Genetics  
Thesis Advisor  
Co-Chair of Committee

Kenneth M Weiss  
Evan Pugh Professor of Biological Anthropology and Genetics  
Co-Chair of Committee

Kateryna Makova  
Assistant Professor of Biology

Hiroshi Akashi  
Assistant Professor of Biology

Esteban J Parra  
Assistant Professor of Molecular Anthropology  
University of Toronto at Mississauga

Richard W Ordway  
Associate Professor of Biology  
Head of the Intercollege Graduate Degree Program in Genetics

*Signatures are on file in the Graduate School
Several complex diseases show population specific differences the causes for which are as yet unknown. Examples of such diseases include obesity and non insulin dependent diabetes which are more prevalent in African American, Indigenous American and Latino populations compared to European Americans. Dementia and Osteoporosis are examples of diseases which are more prevalent in European Americans. It is obvious from these examples that the differences follow broad racial or ethnic categories and because of this correlation, medical research has focused extensively on investigating racial or ethnic differences in disease risk. The concept of ‘Race’ is multifaceted and any simplistic annotation is insufficient, may be even wrong, unless we make attempts to decompose the different aspects of race and use them appropriately in the context of biomedical research. Broadly speaking, race has two distinct components. The biological component referred to as Biogeographical ancestry; and the sociocultural component which reflects the ethnic heritage of an individual. Anthropological and genetic research in the past couple of decades has shown that the concept of typological races is scientifically obsolete yet the idea lives on and is used in biomedical research. Race/Ethnicity is still used as a classifier to establish and highlight population specific differences, while often inter-individual variation within a specific racial/ethnic category is ignored.

Anthropological research has shown that anatomically modern humans originated in Africa, and migrated to different parts of the world. Over the course of time, continental populations were established and for a period, large scale interactions between continental populations were limited. Given the recent and common origin, most genomic regions in all continental populations are very similar. At some loci, however, there has been some change in allele frequencies in the time since the separation of populations. In the past few centuries, there have been large scale migrations, voluntary or forced, that gave rise to populations of “mixed” ancestry. And at those loci where allele frequencies differed between continental populations, this has resulted in long range non random association in the admixed population. Many US residents can trace
their genetic ancestry to more than one continent. The European colonial period that started in the late 1400s brought together in the New World, populations that had been geographically isolated, namely, Europeans, West Africans and Indigenous Americans. The impact of this was two fold, biological, which resulted in genes from different ancestral populations coming together in varying proportions in the different admixed populations and also cultural, in that certain ways of life of one population were adopted by another. Thus, the differences in disease prevalence which appear to follow broad racial categories could have been influenced by both genetic and non genetic factors (including environmental, sociocultural and behavioral factors).

Several resident US populations including European Americans, African Americans and Mexican Americans, which are classified as different races in the biomedical literature have been studied in this thesis. Populations with a recent history of admixture provide a unique opportunity for understanding the genetic bases of many common diseases. If ancestry of alleles at each locus can be established, then it will be possible to correlate the proportion of alleles inherited from an ancestral population to a phenotype and thus establish a genetic basis for the disease. This will help in decomposing the different sources (genetic and non-genetic) that contribute to the phenotype. Thus, estimating ancestry of alleles at a locus is important. Several factors influence the estimate, including nature and number of markers used, the method used for estimating ancestry and the model of admixture that is assumed.

Mitochondrial and Y chromosome segments only provide information regarding paternal or maternal lineage. Autosomal loci on the other hand represent all ancestors of an individual and these are the loci that have been studied in this thesis. The distribution of individual genomic ancestry in different populations has been estimated using several different methods and it is shown that wide variations exist among individuals who self identify as being from the same racial or ethnic category. In addition, significant overlap is seen between two of the racial categories, European Americans and Latinos which are described in this thesis. Modest amounts of geographic variation are also observed within the same racial category. Additional simulation studies have been done to investigate the
effects of markers, model specifications and method choice on the reliability of the estimates of individual genomic ancestry obtained.

Admixture mapping is an approach that uses the ancestry of alleles at a locus in mapping genes for diseases that differ among populations. Several methods have been developed recently to use this approach. This method is suitable for complex diseases since it is unaffected by locus heterogeneity and allelic heterogeneity. In addition unknown gene-environmental interactions further complicate the study of such diseases. Preliminary admixture mapping of hypertension and obesity have been performed and are presented. One locus has been identified that shows strong linkage with obesity in an African American sample.
# TABLE OF CONTENTS

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

LIST OF TABLES ................................................................. xiv

ACKNOWLEDGEMENTS .................................................... xvii

Chapter 1  Measuring and using genomic ancestry ........................................ 1

   Introduction ................................................................. 1
   Understanding ancestry and admixture .................................... 7
   Different aspects of Biogeographical ancestry ............................ 8
   Admixed populations and admixture proportions ..................... 13
   Marker choice for admixture analysis ..................................... 18
     Marker information content ............................................. 19
     Available marker panels: .............................................. 22
   Ancestry component–phenotype correlations ......................... 24
   Methods for detecting stratification ....................................... 28
   Methods for Admixture Mapping: ........................................ 30
   Specific Aims of this project: ............................................ 34
   References ................................................................. 39

Chapter 2  Measurement of admixture proportions and description of admixture
structure in different US populations ........................................ 46

   Introduction: ................................................................. 46
   Materials and Methods: ................................................... 50
     Populations: ............................................................... 50
     Sample collection and genotyping: ....................................... 51
     Estimating individual ancestry: ......................................... 52
     Sample characterization: ............................................... 53
     Detecting Admixture structure: ......................................... 54
   Simulation studies: ......................................................... 54
   Phylogenetic analysis: ..................................................... 56
   Results: ........................................................................ 56
     Allele frequencies and delta ($\delta_c$) levels: ....................... 56
     Estimates of individual ancestry proportions: ..................... 59
     Effect of sample composition on estimates of ancestry ......... 74
     Effect of model choice on ancestry estimates – Using European
       Americans as ancestors: ............................................. 78
   Testing for difference between two Latino Populations .......... 79
   Tests for admixture stratification within samples: ................ 82
   Phylogenetic analysis: ..................................................... 83
   Simulation studies: ......................................................... 84
Appendix B  Perl script for generating individuals with known admixture proportions..................................................................................................................294

Appendix C  Perl script for generating individuals with known admixture proportions and tracking alleles through generations..............................................297

Sim_sample_tag.pl........................................................................................................297

Appendix D  Analyzing individual BioGeographical ancestry and Admixture structure in human populations using 10K markers .........................................302

Appendix E  Perl script for calculating Maximum Likelihood estimates of individual genomic ancestry.........................................................................................313
LIST OF FIGURES

Figure 1.1: Two models for admixture.................................................................3

Figure 1.2: Magnitude of ALD expected under different admixture models ...........4

Figure 1.3: Pedigree representing genealogical ancestry.................................10

Figure 1.4: Distribution of segregation ancestry in a sample of 1000 individuals .....11

Figure 1.5: Correlation between % African ancestry and Melanin index in Puerto Ricans from New York City. Proportional ancestry was measured using 35 AIMs. (Source: Bonilla et al. 2004a).................................................................26

Figure 1.6: Correlation between proportional ancestry and skin pigmentation in Hispanics from San Luis Valley; ancestry estimated using 22 AIMs. (Source: Bonilla et al. 2004b) .................................................................27

Figure 1.7: Correlation between proportional ancestry and skin pigmentation in Mexicans from Tlapa, Mexico measured using 29 AIMs (Source: Bonilla 2003)..............................................................................................................28

Figure 2.1: Distribution of Individual ancestry proportions in 652 European Americans from Connecticut (EACT) using Maximum Likelihood. WAF: West African, IA: Indigenous American; EU: European ......................................................60

Figure 2.2: Distribution of Individual ancestry proportions in 228 African Americans from Connecticut (AACT) using Maximum Likelihood. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry.................................................................62

Figure 2.3: Distribution of Individual ancestry proportions in 102 Latinos from Connecticut (LCT) using Maximum Likelihood. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry.................................64

Figure 2.4: Distribution of Individual ancestry proportions in 89 Latinos from California (LCA) using Maximum Likelihood. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry.................................65

Figure 2.5: Distribution of Individual ancestry proportions in 652 European Americans from Connecticut (EACT) using STRUCTURE. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry....67

Figure 2.6: Distribution of Individual ancestry proportions in 228 African Americans from Connecticut (AACT) using STRUCTURE. IA: Indigenous
American ancestry, WAF: West African ancestry and EU: European ancestry

Figure 2.7: Distribution of Individual ancestry proportions in 102 Latinos from Connecticut (LCT) using STRUCTURE - IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry. Proportional EU

Figure 2.8: Distribution of Individual ancestry proportions in 89 Latinos from California (LCA) using STRUCTURE. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry

Figure 2.9: Correlation between estimates obtained with ML and STRUCTURE

Figure 2.10: Comparing effect of large sample size (left panels) to small sample size (right panels) in EACT samples

Figure 2.11: Comparing effect of large sample size (left panels) to small sample size (right panels) in AACT samples

Figure 2.12: Cumulative Fraction plots from K-S tests comparing the two Latino populations

Figure 2.13: Phylogenetic tree of all populations constructed using the Neighbor-Joining method implemented in MEGA 2

Figure 2.14: Cumulative WAF ancestry distribution in simulated samples using ML (shown in top panel) and using STRUCTURE (shown in bottom panel)

Figure 3.1: Distribution of ancestry estimates in African American samples from Bogalusa, LA (N=176; dots) and Philadelphia, PA (N=202, blank circles).

Figure 3.2: Histogram of distribution of ancestry proportions in an African-American metapopulation. Proportional ancestry is shown on X axis and number of individuals on the Y axis

Figure 3.3: Distribution of proportional ancestry estimates in the African-American metapopulation using ADMIXMAP. Solid line indicates the theoretical Dirichlet distribution, histogram indicates observed ancestry proportions

Figure 4.1: Distribution of composite delta between all pair wise populations in the ancestral population. δ is highest between Africans (AF) and Indigenous Americans (IA) and lowest between Europeans (EU) and Indigenous Americans
Figure 4.2: Distribution of individual genomic ancestry estimates in 243 European Americans from GenNET estimated using ML....................................................159

Figure 4.3: Distribution of individual genomic ancestry estimates in 425 European Americans from GENOA estimated using ML ....................................................160

Figure 4.4: Distribution of individual genomic ancestry estimates in 607 European Americans from HyperGEN estimated using ML ................................................161

Figure 4.5: Distribution of individual genomic ancestry estimates in 243 European Americans from GenNET estimated using STRUCTURE. .................................162

Figure 4.6: Distribution of individual genomic ancestry estimates in 425 European Americans from GENOA estimated using STRUCTURE...................................163

Figure 4.7: Distribution of individual genomic ancestry estimates in 607 European Americans from HyperGEN estimated using ML ................................................164

Figure 4.8: Plot of all 242 individuals within GenNET, including 28 with high levels of Indigenous American ancestry, who were not used for stratification or phenotype analysis. .................................................................165

Figure 4.9: Distribution of ancestry estimates and proportional ancestry in 278 African Americans from GenNET using ML.......................................................169

Figure 4.10: Distribution of ancestry estimates and proportional ancestry in 542 African Americans from GENOA estimated using ML. ......................................170

Figure 4.11: Distribution of ancestry estimates and proportional ancestry in 698 African Americans from HyperGEN using ML. ..................................................171

Figure 4.12: Distribution of ancestry estimates and proportional ancestry in 278 African Americans from GenNET using STRUCTURE......................................172

Figure 4.13: Distribution of ancestry estimates and proportional ancestry in 542 African Americans from GENOA using STRUCTURE. .................................173

Figure 4.14: Distribution of ancestry estimates and proportional ancestry in 698 African Americans from HyperGEN using STRUCTURE..................................174

Figure 4.15: Distribution of ancestry estimates and proportional ancestry in 413 Mexican Americans from GENOA using ML......................................................177

Figure 4.16: Distribution of ancestry estimates and proportional ancestry in 413 Mexican Americans from GENOA using STRUCTURE. .................................178
Figure 4.17: Plot of 50 European-American individuals with highest Indigenous American ancestry (blank circles) from the GENOA network and 50 Mexican American individuals with lowest Indigenous American ancestry (dots) from the combined European-American sample.................................179

Figure 4.18: KS test comparison percentile plots comparing expected ancestry (solid line) against estimated ancestry (broken lines) using ML (left side panels) and STRUCTURE (right side panels). A) Plots comparing proportional West African ancestry, B) Plots comparing European ancestry, C) Plots comparing Indigenous American ancestry .............................................190

Figure 5.1: Triangle plots of ancestry estimates in simulated samples using the balanced panel of 102 AIMs. ..........................................................................................238

Figure 5.2: Triangle plots of ancestry estimates in simulated samples using the unbalanced panel of 100 AIMs.................................................................240

Figure 5.3: Cumulative Percentile plots of distribution of ancestry for each axis in MLE-UB and MLE-B......................................................................................243

Figure 5.4: Cumulative Percentile plots of distribution of ancestry for each axis in NP-S-UB and NP-S-B...................................................................................245

Figure 5.5: Cumulative Percentile plots of distribution of ancestry for each axis in NP-I-UB and NP-I-B......................................................................................246

Figure 5.6: Cumulative Percentile plots of distribution of ancestry for each axis in NF-I-UB and NF-I-B......................................................................................248

Figure 5.7: Cumulative percentile plots of distribution of ancestry for each axis in NF-S-UB and NF-S-B......................................................................................249

Figure 5.8: Cumulative fraction plots for option F-S-UB: Using Popflag option, same alpha option and unbalanced marker panel...........................................251
LIST OF TABLES

Table 1.1: Diseases with possible genetic components based on ethnic differences in disease rates and hence amenable to admixture mapping………………………………6

Table 1.2: Proportion of European ancestry in different African American populations and one African Caribbean population ………………………………………..15

Table 1.3: Admixture proportions in different Latino Populations …………………………..17

Table 1.4: Measures of Marker ancestry information content……………………………..21

Table 2.1: Allele frequency differences (δc) in different populations ……………………58

Table 2.2: Mean estimates of individual genomic ancestry using ML and STRUCTURE…………………………………………………………………………………..59

Table 2.3: Correlation between estimates obtained with ML and STRUCTURE…..73

Table 2.4: Effect of sample size on Ancestry estimates in EACT using STRUCTURE…………………………………………………………………………………..75

Table 2.5: Effect of sample size on Ancestry estimates in AACT using STRUCTURE…………………………………………………………………………………..75

Table 2.6: Comparing ancestry estimates in two Latino populations…………………80

Table 2.7: Results of individual ancestry correlation test…………………………………82

Table 2.8: Estimated ancestry in simulated sample of unadmixed individuals………86

Table 2.9: Ancestry composition in simulated samples………………………………………89

Table 2.10: Mean estimates of ancestry obtained with MLE and STRUCTURE in simulated sample…………………………………………………………………..89

Table 2.11: Correlation in estimates of simulated individuals for ML and STRUCTURE………………………………………………………………………..90

Table 3.1: Locus ID, polymorphism, chromosomal position, allele frequency in different populations, allele frequency differences, tests of Hardy Weinberg equilibrium………………………………………………………………………..113

Table 3.2: Characteristics of combined sample………………………………………………117

Table 3.3: Characteristics of sample from each location……………………………………118
Table 3.4: Sex specific estimates of phenotypes in total sample..............................118
Table 3.5: Sex specific distributions of phenotypes in different locations............119
Table 3.6: Associations between genotypes and phenotypes.................................124
Table 3.7: Summary of posterior parameters estimated using ADMIXMAP..........127

Table 4.1: Mean ± Standard Deviation of proportional ancestry estimates in FBPP samples using Maximum Likelihood..........................................................157
Table 4.2: Mean ± Standard Deviation of proportional ancestry estimates in FBPP samples using STRUCTURE ..........................................................158
Table 4.3: Mean ± Standard Deviation of proportional ancestry estimates in European Americans from different field centers within HyperGEN samples ....167
Table 4.4: Correlation between proportional ancestry estimates obtained with ML and STRUCTURE..........................................................175
Table 4.5: Mean ± Standard Deviation of proportional ancestry estimates in African Americans from different field centers within HyperGEN samples ......176
Table 4.6: Results of Individual Ancestry Correlation Test using Maximum Likelihood..........................................................................................180
Table 4.7: Results of Individual Ancestry Correlation Test using STRUCTURE ......181
Table 4.8: Phenotypic characteristics of FBPP samples...........................................183
Table 4.9: Associations between ancestry and phenotypes in FBPP populations ......185
Table 4.10: Mean ± Standard Deviation of proportional ancestry estimates in simulated samples with 100% ancestry from one population. .......................187
Table 4.11: Distribution of ancestry proportions in simulated sample...............188
Table 4.12: Mean ± Standard Deviation of proportional ancestry estimates in simulated sample of mixed ancestry .........................................................189
Table 4.13: Correlation between expected and observed estimates of ancestry......189
Table 5.1: Distribution of genealogical ancestry in sample of simulated individuals........................................................................................................223
Table 5.2: Balanced SNP Marker Panel selected from Affymetrix 10k WGA Chip...........................................................................................................228
Table 5.3: Unbalanced SNP Marker Panel selected from Affymetrix 10k WGA Chip

Table 5.4: Results for Balanced Panel of markers: Mean (Variance) for each ancestry axis, MSE, Correlations between expected and observed estimates of ancestry and Results of t-test and KS tests

Table 5.5: Results for Unbalanced Panel of markers: Mean (Variance) for each ancestry axis, MSE, Correlations between expected and observed estimates of ancestry and Results of t-test and KS tests
ACKNOWLEDGEMENTS

The two people who are most responsible for my success are my parents who made me believe that I could achieve my goals if I worked hard, and let me do things my way. Their faith and support have given me strength and confidence in myself. My life in the US would have been different if it had not been for the love and support of my husband Sid. I want to thank Sid for going through this journey with me, for helping me stay focused in all that I had to do to complete my dissertation.

None of this would have happened of course if I did not have Mark Shriver as my advisor. He introduced me to a whole new area of research and constantly challenged me to do my best. He also appears to have had extraordinary faith in my abilities, than any other advisor would. I thank him for giving me the freedom to explore, and sustenance to grow as a scientist.

I want to thank Dr. Kenneth Weiss, whose insight into the conundrum of complex disease research has made me look at everything I do critically and has taught me to look between the lines of otherwise glamorous research. I have always feared his ‘Now, wait a minute….’ to interrupt my talks but in the end these interruptions have helped me clarify many important concepts.

I want to thank Anne Buchanan who has been a very good friend and for being the person to whom I could run to whenever I needed encouragement and consolation.

I am grateful to Esteban Parra and Carolina Bonilla, former lab seniors, who I have bothered often and talking to them helped me understand concepts. Special thanks to Vibhor Sonpar who was marvelous with programming. Interacting with my lab mates Heather Norton, Tamiko Brown, Abby Bigham, Sonia Dios, Paula Goetting, Marc Bauchet and Xianyun Mao has been very exciting and stimulating and has enriched my graduate experience at Penn State.

My time in State College would not have been as fun if it were not for my dear friends Sanjukta and Amitayu who rejoiced with me during good times and stood by me during not so good times.
Chapter 1

Measuring and using genomic ancestry

Introduction

The European colonial period that started in the late 1400s brought together in the New World populations that had been geographically isolated for tens of thousands of years, namely, Europeans, West Africans and Indigenous Americans. These events led to the creation of many “admixed” populations, members of which can trace their genetic ancestry to more than one continent. Given the recent and common origin of all human populations, this admixture had only a small average effect on the gene pools of these new populations. In other words, for most genomic regions, the pre-colonial (or parental) populations had similar allele frequencies and, at these, admixture was of little consequence. At some other loci, however, there had been some change in allele frequencies in the time since the separation of parental populations and it is at these loci where admixture has had an important effect. The process of admixture creates non-random allelic associations or linkage disequilibrium (LD) between all loci, linked and unlinked that have different allele frequencies in the parental populations. Since populations like African Americans, African Caribbeans and Mexican Americans were formed in the recent past, allelic associations in these populations that were created by admixture extend over large distances. Such populations represent a useful resource for mapping complex-disease genes by using this long-range admixture linkage disequilibrium (ALD) (Chakraborty and Weiss 1988; Risch 1992; Stephens et al. 1994), which require fewer markers to screen the genome than other populations or approaches. Understanding the genetic consequences of admixture is important because it can be both a confounding factor and a source of statistical power in gene identification studies.
The magnitude of ALD in an admixed population depends on the allele frequency differential between the parental populations, the level of admixture, the admixture dynamics, the time since admixture and the recombination rate between the loci (Chakraborty and Weiss 1988; Briscoe et al. 1994; Stephens et al. 1994). While ALD between unlinked markers decays rapidly (within two to four generations), ALD between linked markers decays more slowly. The exponential decrease in ALD with genetic distance facilitates the differentiation of ALD that is high between markers that are close together and genetically linked, from ALD generated at unlinked loci. Thus, if the parental populations differ in a trait or disease due to different frequencies of risk alleles, it should be possible to identify the loci containing these alleles using admixture mapping (AM). (Chakraborty and Weiss 1988; McKeigue 2000; McKeigue et al. 2000; Shriver et al. 2003).

Two models of admixture dynamics have been described which represent the extremes of the process by which an admixed population is formed: the continuous gene flow (CGF) model and the hybrid isolation (HI) model (Long 1991; Pfaff et al. 2001). As shown in Figure 1.1, in the HI model, admixture occurs in a single generation without further contribution from either parental population, hence, ALD is generated in a single generation and gradually decays in successive generations through independent assortment and recombination between loci. Few false-positive results are thus expected in an association study under the HI model. Alternatively, the CGF model (Figure 1.1) represents the situation at the other end of the spectrum. In this case admixture occurs in each generation, with contributions from one (or all) of the parental populations into the admixed population. In reality, populations may have an admixture history where several scenarios along the admixture model spectrum have played a part. ALD under the CGF model increases in each generation, since new admixture is constantly occurring. A point will be reached, however (when the admixture proportion ~ 0.5), where continued admixture will actually decrease the ALD, since added gene flow will result in the conversion of the admixed population into the introgressing parental population. Figure 1.2 shows the amount of ALD expected under these two models for linked and unlinked loci. For both models, association between markers is inversely correlated with
the genetic distance between them. Simulation studies have shown that populations that have a demographic history more consistent with the CGF model of admixture retain ALD over larger chromosomal regions and show significant associations between unlinked marker loci (Pfaff et al. 2001). While associations between unlinked markers could potentially lead to false-positives, conditioning upon parental admixture allows the distinction between associations arising due to true linkage and those due to admixture stratification allowing for tests of linkage in the presence of stratification (McKeigue 1998; McKeigue 2000; McKeigue et al. 2000).

The model on the left follows Hybrid Isolation, where ancestral populations admix in a single generation, G1. No further contribution is received from ancestral populations. The model on the right shows Continuous gene flow where the admixed population is created in generation G1. In successive generations one or all ancestral populations may continue to contribute to the admixed population.

Figure 1.1: Two models for admixture.
The amount of admixture linkage disequilibrium (ALD) expected under the continuous gene flow (CGF) and hybrid isolation (HI) models of admixture for unlinked loci and loci linked at 5 cM. The results shown are for two loci with $\delta$ 0.54 and 0.49, and with 50 per cent admixture in the first generation for the HI model and 1.9 per cent admixture for 36 generations under the CGF model (equivalent to 50 per cent total). ALD under the HI model decreases for both linked and unlinked loci, whereas ALD under the CGF model for both linked and unlinked loci increases initially and then decreases (from Pfaff et al., 2001)

Figure 1.2: Magnitude of ALD expected under different admixture models
There are several ways in which admixture can be an important resource in the elucidation of genetic factors that contribute to the risk of common disease. Common diseases often have environmental components to their risk, and the clinical phenotype results from currently unknown interactions between environmental factors and underlying genotypes. Decomposing the sources of variation is thus important in order to accurately understand the etiology of the trait. It may be possible to distinguish between the genetic and environmental explanations for ethnic differences in disease risk (and investigating the mode of inheritance), by studying the relationship of disease risk to individual admixture (McKeigue 1997; Fernandez et al. 2003; Molokhia et al. 2003; Shriver et al. 2003; Bonilla et al. 2004). For example, recent studies have demonstrated a strong relationship between proportional West African ancestry and the risk of systemic lupus erythematosus in admixed populations in Trinidad (Molokhia et al. 2003). Several common diseases (e.g. hypertension, diabetes, obesity, prostate cancer and osteoporosis) have differences in risk among population groups which has been speculated, to result in part, from genetic causes as shown in Table 1.1. In situations where these differences have a genetic basis, testing for locus ancestry conditioning on parental admixture can help map genes underlying these differences. This approach has a greater statistical power than family linkage studies for mapping polygenic traits (McKeigue 1997; McKeigue 1998; Shriver et al. 2003). Estimates of BioGeographical Ancestry (BGA), the proportional ancestry from different ancestral populations in an individual, can be used in conjunction with measured environmental effects for investigating the relative roles of environmental and inherited risks underlying complex traits (Fernandez et al. 2003; Gower et al. 2003; Molokhia et al. 2003; Shriver et al. 2003; Bonilla et al. 2004b; Bonilla et al. 2004a). It is important to recognize that associations between individual admixture and disease risk might reflect correlations between BGA and socio-cultural variables and exposures. For example, hypothetically, if BGA and years of education were to be correlated, hypertension might be correlated with BGA, even if the underlying risk factor were years of education. Alternatively, a correlation between BGA and phenotype may correctly reflect a fundamental effect of BGA on the phenotype.
Table 1.1: Diseases with possible genetic components based on ethnic differences in disease rates and hence amenable to admixture mapping

<table>
<thead>
<tr>
<th>Disease</th>
<th>High Risk Group</th>
<th>Low Risk Group</th>
<th>Relative Risk Ratio</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-insulin dependent diabetes (NIDDM)</td>
<td>South Asians, West Africans, Peninsular Arabs, Pacific Islanders and Indigenous Americans</td>
<td>Europeans</td>
<td>4:7</td>
<td>(Martinez 1993; Songer and Zimmet 1995)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>African Americans, West Africans</td>
<td>Europeans</td>
<td>2:3</td>
<td>(Gaines and Burke 1995; Douglas et al. 1996)</td>
</tr>
<tr>
<td>End-stage renal disease</td>
<td>Indigenous Americans and African populations</td>
<td>Europeans</td>
<td>N/A</td>
<td>(Ferguson and Morrissey 1993)</td>
</tr>
<tr>
<td>Dementia</td>
<td>Europeans</td>
<td>African Americans, Latino Americans</td>
<td>N/A</td>
<td>(Hargrave et al. 2000)</td>
</tr>
<tr>
<td>Autoimmune diseases:</td>
<td>West Africans</td>
<td>Europeans</td>
<td>N/A</td>
<td>(Molokhia and McKeigue 2000)</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>Indigenous Americans</td>
<td>Europeans</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skin cancer</td>
<td>Europeans</td>
<td>European Americans, Chinese, Japanese</td>
<td>N/A</td>
<td>(Shimizu et al. 1985; Schwartz and Swanson 1997; Hoffman et al. 2001; Boni et al. 2002)</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>Africans</td>
<td>Europeans</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>Africans and African Americans</td>
<td>European Americans, Chinese, Japanese</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>Europeans</td>
<td>Chinese, Japanese, African Americans, Turkmen, Uzbeks, Native Siberians, New Zealand Maoris</td>
<td>N/A</td>
<td>(Rosati 2001)</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>European Americans</td>
<td>African Americans</td>
<td>N/A</td>
<td>(Bohannon 1999)</td>
</tr>
</tbody>
</table>
Understanding ancestry and admixture

As defined in the previous section, individual BGA refers to the proportional ancestry from different ancestral populations. To obtain a perfect estimate of BGA of an individual one would have to sequence the entire genome of a person and determine the population of origin of each locus together with sequencing many individuals from the precise ancestral populations. However, when we estimate individual BGA, we generally use a subset of loci from the genome and make reasonable assumptions about ancestral populations. One reason for doing this is that obtaining reliable estimates of BGA requires ascertaining ancestral allele frequencies and/or marker genotypes from modern descendants of appropriate ancestral populations. Without estimates of ancestral allele frequencies and/or genotypes it is difficult to accurately infer locus ancestry in an admixed individual. Thus, there is a practical consideration of how many and which loci can be sampled for reliable BGA estimates. In practice we use specific loci that can be considered as ancestry informative markers (AIMs) in that they have large allele frequency differences between parental populations. This allows us to make some conclusions regarding the probability of an allele at a locus being inherited from one ancestral population vs. another. However, most AIMs are not fixed among ancestral populations and so we cannot rule out the chance that the ancestry of the allele could in reality, be from another population. For instance, presence of the duffy null (FY*0 - rs2814778) allele in an individual indicates sub-Saharan African ancestry at that locus since FY*0 is fixed in many sub-Saharan populations and implies at least one sub-Saharan African ancestor in that individuals’ genealogy. However, observing the T allele at the LPL locus (rs285) only suggests a higher probability of that allele being inherited from an African ancestor (allele frequency in West Africans ~ 98%) as opposed to inheriting the allele from a European ancestor (allele frequency in Europeans ~ 50%) or a Indigenous American ancestor (allele frequency ~ 45% in Indigenous Americans). It is standard practice to increase confidence in BGA estimates by combining information across multiple loci. Groups of closely linked loci can be analyzed as haplotypes, as a
measure to generate regional genomic BGA estimates that are more reliable. In admixed individuals ancestry varies across different loci or different genomic segments. Genomic ancestry for such individuals is estimated as a weighted average of ancestry across multiple loci. This is necessary; since estimates obtained using single locus specific ancestry will vary between loci in an individual. For instance, we may observe the FY*0 allele at a locus and conclude African ancestry for an individual, but if we observe the MID 575 (rs140864) insertion polymorphism, which is also on the same chromosome as FY*0, then we would have to conclude greater chance of European ancestry for the same individual, at that locus. Instead, it is possible to make genome wide inferences of admixture proportions by combining information across multiple loci. Several methods have been developed for this purpose and are described in a later section.

**Different aspects of Biogeographical ancestry**

While measuring admixture proportions the issue of reliability of estimates obtained is important. In this context, it is important to discuss certain terms and explain how these have been used in this study. Biogeographical ancestry broadly refers to three related concepts: First, there is **Genealogical ancestry** that refers to the proportion of ancestors in an individual’s genealogy that belong to each of the ancestral populations. Thus an individual who has two grandparents from one ancestral population and two from another ancestral population has 50% genealogical ancestry from each population. However, due to the process of segregation the proportion of alleles that are actually inherited from an ancestral gene pool may differ from the genealogical ancestry. This is referred to as **Segregation ancestry**. Lastly, there is **Genomic ancestry** or the proportional ancestry that is measured using different markers that are informative for ancestry. Admixture refers to the process by which individuals from different populations come together to create a new population. These concepts are further elaborated using Figure 1.3. This figure represents a pedigree of 15 individuals. In this figure individuals from generation I, viz. I₁, I₃, I₅, I₆ and I₈ are from ancestral population 1 (Pop1) (Blank
symbols), individuals I\textsubscript{2} and I\textsubscript{4} are from ancestral population 2 (Pop2) (Blue symbols) and individual I\textsubscript{7} is from ancestral population 3 (Pop3) (Grey symbols). Admixture events occur in the first generation, and individuals I\textsubscript{9}, I\textsubscript{10} and I\textsubscript{12} are all admixed individuals in the second generation. In generation II individuals I\textsubscript{9} and I\textsubscript{10} both have 0.5 genealogical ancestry from Pop1 and Pop2 respectively, since they both have one parent each from Pop1 and one from Pop2. Individual I\textsubscript{12} has 0.5 genealogical ancestry from Pop1 and also from Pop3. In generation III, individual I\textsubscript{13} who inherits half of his genome from each parent is expected to have 0.5 ancestry from Pop1 and Pop2, since the genealogical ancestry in an individual is expected to be the average of that of the parents. Following a similar argument, individual I\textsubscript{14} is expected to have 0.25 ancestry from Pop3 and 0.75 from Pop1. Individuals I\textsubscript{15} is expected to have 0.625 ancestry from Pop1, 0.25 ancestry from Pop2 and 0.125 ancestry from Pop3. In each case, the estimates of BGA indicated above refer to that individuals’ genealogical ancestry, since these estimates refer to the number of ancestors of that individual who are from one particular ancestral population. However, the genealogical ancestry is a theoretical expectation that may or may not be realized. This is because of independent assortment and segregation of alleles at unlinked loci in each generation. Only when an individual has two parents who are from two different ancestral populations and who themselves do not have a recent history of admixture, will genealogical ancestry be equal to segregation ancestry. However, for all successive generations there will be some variation in segregation ancestry.
Individuals in generation I all have 100% ancestry from one population. Admixed individuals are seen from generation II onwards. Individuals $I_9$, $I_{10}$ and $I_{12}$ have exactly 50% genealogical and also 50% segregation ancestry. From generation III onwards segregation ancestry will differ from genealogical ancestry.
If we had exactly 1000 such pedigrees and we were able to accurately measure the segregation ancestry for 1000 individuals (Chakraborty and Weiss 1986) then the genealogical ancestry for all these individuals would be exactly the same, although their segregation ancestries would differ. The average of the segregation ancestry estimates of all 1000 individuals would however approach the expected genealogical ancestry. Figure 1.4 shows the distribution of segregation ancestry in sample of 1000 such individuals.

Figure 1.4: Distribution of segregation ancestry in a sample of 1000 individuals
As seen in Figure 1.4 even within four generations segregation ancestry may show significant deviation from genealogical ancestry. In the above example a panel of 90 hypothetical markers was used, where alternate alleles are fixed in different populations, such that 30 markers are each absolutely informative for one ancestral population (Details of the simulation are provided in Chapter 4 of this thesis). In the above example, proportional segregation ancestry from Pop1 ranges from 0.55 to 0.75, with the mean being 0.63. Ranges of variation are also observed in proportional segregation ancestry from Pop2 and Pop3. Only if there were an infinite number of segregating sites in the genome, would genealogical ancestry be equal to segregation ancestry. However, the genome is finite and the numbers of segregating sites are limited. Hence, chances of segregation ancestry differing from genealogical ancestry are always going to be high.

The aspect of BGA that is measured using markers is referred to as genomic ancestry. Measures of genomic ancestry, which are calculated from average admixture proportions across the genome, depend on the marker information content of the AIMs (how good the AIMs are in distinguishing between populations), the number of AIMs used and the genomic coverage of the AIMs (how evenly the AIMs are spread over the genome). The error in genomic ancestry estimates will reduce and gradually approach zero as the number and coverage of AIMs across the genome increases. Genomic ancestry $y$ would numerically be the equal to both segregation ancestry and genealogical ancestry if and only if there were an infinite number of segregating sites and all sites (loci) were perfectly informative for ancestry (i.e. alternate alleles fixed in each parental population). However, the genome is of finite size and there are few loci at which there has been an evolutionary effect that has led to fixation of an allele in a population. Thus, not only is there variation due to differences between genealogical and segregation ancestry, additional variation is introduced by the use of the particular markers that are used for measuring genomic ancestry. Another way to think about the difference is that, at each generation the genomic ancestry changes from the genealogical ancestry due to segregation.
Besides, genomic ancestry is not distributed evenly across the chromosomes of the parents, although the expected levels of ancestry are the same throughout the genome. Both genomic and genealogical ancestry for a person reflects their “true” ancestry. Genealogical ancestry for a person is established and unchangeable, given that person’s individual history. Genomic ancestry is also fixed for a person and depends on the exact alleles inherited by an individual. Deciding on which description of ancestry should be used depends on specific questions that are being asked. For assessing ancestry-phenotype associations and admixture mapping, a measure of genomic ancestry is required. Methods have been developed to measure genomic ancestries in individuals and sometimes these measured estimates of genomic ancestry are used to make inferences about genealogical ancestry of a person. However, since it is often difficult to ascertain exactly which specific alleles from a parent will be inherited by an offspring, it is difficult to quantify the variation in genomic ancestry from the expected genealogical ancestry measure and to then use this estimate of genomic ancestry as an “expected” variable.

In practice, the focus of this thesis has been on trying to estimate genomic ancestry proportions and characterize the distribution of ancestry in several populations and to make inferences about the genetic basis of stratification in the admixed populations. Measures of segregation ancestry has been used as the statistical expectation in the simulation studies performed in this thesis, to understand how different programs perform in measuring genomic ancestry and to make some meaningful inferences regarding observed ancestry distributions.

**Admixed populations and admixture proportions**

Since the amount of ALD created is proportional to the level of admixture in a population, it is important to briefly review studies on admixture levels across populations. Those populations that are likely to be useful for admixture studies include African Americans, Mexican Americans, Cubans and Puerto Ricans in the USA, African
Caribbeans, various Latin American populations, various groups in Central and South America and the Caribbean islands, Anglo Indians in India and ‘colored’ populations of South Africa. Various statistical approaches have been used to estimate admixture proportions (both in groups and in individuals) in these populations and have been reviewed in detail by (Chakraborty 1986). These methods include a least squares method, a weighted least squares method (Elston 1971; Long and Smouse 1983; Long 1991), likelihood methods (Long 1991; Chikhi et al. 2001; Tang et al. 2005), and Bayesian methods (McKeigue 2000; Pritchard et al. 2000; Falush et al. 2003; Hoggart et al. 2003).

Admixture proportions in African-American populations are highly variable across the USA, which is likely to reflect local variation in the demographic histories and social norms. Table 1.2 lists several studies in which admixture proportions in African Americans and two African Caribbean populations have been measured. Shown are mean estimates of admixture proportions. This is a well-studied metapopulation with substantial European and West African contributions and a smaller Indigenous American contribution (Chakraborty et al. 1992; Parra et al. 1998; Destro-Bisol et al. 1999; Shriver et al. 2003). A survey of current literature (as shown in Table 1.2) indicates that European admixture ranges from 0.035 in the Gullah Sea Islanders of South Carolina (Parra et al. 1998) to 0.31 in Baltimore (Glass and Li 1953).
Table 1.2: Proportion of European ancestry in different African American populations and one African Caribbean population

<table>
<thead>
<tr>
<th>Location</th>
<th>European</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pittsburgh</td>
<td>0.25 ± 0.02</td>
<td>(Chakraborty et al. 1992)</td>
</tr>
<tr>
<td>Maywood, IL</td>
<td>0.18 ± 0.01</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>Detroit</td>
<td>0.16 ± 0.02</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>New York</td>
<td>0.19 ± 0.02</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>Philadelphia 1</td>
<td>0.12 ± 0.01</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>Philadelphia 2</td>
<td>0.13 ± 0.01</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>0.20 ± 0.01</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>Baltimore</td>
<td>0.15 ± 0.02</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>Charleston, SC</td>
<td>0.12 ± 0.01</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>New Orleans</td>
<td>0.22 ± 0.01</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>Houston</td>
<td>0.16 ± 0.01</td>
<td>(Parra et al. 1998)</td>
</tr>
<tr>
<td>Detroit</td>
<td>0.26</td>
<td>(Reed 1969)</td>
</tr>
<tr>
<td>New York</td>
<td>0.18</td>
<td>(Reed 1969)</td>
</tr>
<tr>
<td>Baltimore</td>
<td>0.31</td>
<td>(Glass and Li 1953)</td>
</tr>
<tr>
<td>Claxton, Ga</td>
<td>0.14 ± 0.05</td>
<td>(Long 1991)</td>
</tr>
<tr>
<td>Sapelo Island, Ga</td>
<td>0.06 ± 0.05</td>
<td>(Long 1991)</td>
</tr>
<tr>
<td>Gullah Island, SC</td>
<td>0.03</td>
<td>(Parra et al. 2001)</td>
</tr>
<tr>
<td>Columbia, SC</td>
<td>0.17</td>
<td>(Parra et al. 2001)</td>
</tr>
<tr>
<td>Chicago</td>
<td>0.26 ± 0.02</td>
<td>(Destro-Bisol et al. 1999)</td>
</tr>
<tr>
<td>Washington DC</td>
<td>0.19 ± 0.02</td>
<td>(Shriver et al. 2003)*</td>
</tr>
<tr>
<td>CHS **</td>
<td>0.21 ± 0.01</td>
<td>(Reiner et al. 2005)*</td>
</tr>
<tr>
<td>African Caribbeans, UK</td>
<td>0.1 ± 0.14</td>
<td>(Shriver et al. 2003)*</td>
</tr>
</tbody>
</table>

*Measured admixture under a three way model with proportional European, West African and Indigenous American ancestral populations. Proportional Indigenous American ancestry for samples from Washington DC, UK and CHS are 0.027 ± 0.014, 0.019 ± 0.013 and 0.027 ± 0.016 respectively.

** CHS: Cardio Vascular Heart Study, population samples from Pittsburgh, PA, Winston-Salem, NC, Washington County, MD, Sacramento, CA.
US Latinos form a complex socio-political conglomerate including Puerto Ricans, Cubans, Spanish Americans, Mexican Americans, and Central and South Americans living in the US. In many of these Latino populations the proportional contributions from parental Europeans are estimated to be the largest, followed by a substantial Indigenous American ancestry and varying amounts of West African ancestry (Hanis et al. 1986; Long 1991; Bonilla et al. 2004a; Bonilla et al. 2004b), while in other samples proportional West African (Crawford et al. 1981; Bortolini et al. 1995) or proportional Indigenous American (Lisker and Babinsky 1986; Lisker et al. 1986b) may be the highest. Table 1.3 lists several studies on Latino populations along with mean admixture proportions observed in these samples. As with African-American populations, there is substantial variation across populations. From these results, it is evident that, when studying any new admixed population sample, it is important to estimate the proportional contributions in the population under consideration and not to rely on previously obtained estimates from a similar population. Additionally, it is instructive to have information on the levels of stratification related to admixture that are present in the population under consideration (Pfaff et al. 2001), since undetected stratification will cause confounding in association studies. Besides, this stratification is also the basis for ancestry-phenotype associations and can be used to investigate genetic basis underlying a phenotype.
<table>
<thead>
<tr>
<th>Location (Population)</th>
<th>European</th>
<th>African</th>
<th>Indigenous American</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arizona (Mexican Americans)</td>
<td>0.68 ± 0.05</td>
<td>3 ± 2</td>
<td>29 ± 4</td>
<td>(Long 1991)</td>
</tr>
<tr>
<td>Mexican Americans, TX</td>
<td>0.65</td>
<td></td>
<td>0.35</td>
<td>(Hantis et al. 1986)</td>
</tr>
<tr>
<td>Gila River Indian Community, AR</td>
<td>0.305</td>
<td>0.023</td>
<td>0.671</td>
<td>(Williams et al. 2000)</td>
</tr>
<tr>
<td>San Luis Valley, CO</td>
<td>0.63 ± 0.02</td>
<td>0.03 ± 0.01</td>
<td>0.34 ± 0.02</td>
<td>(Bonilla et al. 2004b)</td>
</tr>
<tr>
<td>Puerto Rican, NYC</td>
<td>0.53 ± 0.03</td>
<td>0.29 ± 0.02</td>
<td>0.17 ± 0.02</td>
<td>(Bonilla et al. 2004a)</td>
</tr>
<tr>
<td>Mexican-American, CA</td>
<td>0.60</td>
<td></td>
<td>0.4</td>
<td>(Collins-Schramm et al. 2004)</td>
</tr>
<tr>
<td>Santiago, Chile</td>
<td>0.57</td>
<td></td>
<td>0.43</td>
<td>Reviewed by (Sans 2000)</td>
</tr>
<tr>
<td>Mexico City, Mexico</td>
<td>0.41</td>
<td>0.03</td>
<td>0.056</td>
<td>(Lisker et al. 1986b)</td>
</tr>
<tr>
<td>La Plata, Argentina</td>
<td>0.63</td>
<td>0.07</td>
<td>0.30</td>
<td>Reviewed by (Sans 2000)</td>
</tr>
<tr>
<td>Montevideo, Uruguay</td>
<td>0.9</td>
<td>0.09</td>
<td>0.01</td>
<td>(Sans et al. 1997)</td>
</tr>
<tr>
<td>Melo, Uruguay (Blacks)</td>
<td>0.35</td>
<td>0.52</td>
<td>0.13</td>
<td>(Sans et al. 2002)</td>
</tr>
<tr>
<td>Paredao, Brazil</td>
<td>0.18</td>
<td>0.79</td>
<td>0.03</td>
<td>(Bortolini et al. 1995)</td>
</tr>
<tr>
<td>Cameta, Brazil</td>
<td>0.18</td>
<td>0.48</td>
<td>0.34</td>
<td>(Bortolini et al. 1995)</td>
</tr>
<tr>
<td>Punta Gorda, Belize</td>
<td>0.05</td>
<td>0.71</td>
<td>0.24</td>
<td>Reviewed by (Sans 2000)</td>
</tr>
<tr>
<td>Livingston, Guatemala</td>
<td>0.01</td>
<td>0.7</td>
<td>0.29</td>
<td>(Crawford et al. 1981)</td>
</tr>
</tbody>
</table>
Marker choice for admixture analysis

Admixture-based methods rely on using suitable markers and estimates of allele frequencies from appropriately identified parental populations. Since ALD is fairly new and extends over larger distances, fewer markers are required for AM studies. Markers informative for ancestry have been used in several contexts and have been referred to as ‘ideal,’ (Reed 1973) ‘private’ (Neel 1974) and ‘unique’ (Chakraborty et al. 1991). Informativeness of such markers can be measured as the allele frequency differential ($\delta$), which is the absolute value of the difference of a particular allele between populations (Chakraborty and Weiss 1986; Dean et al. 1994) (detailed in the next section). Microsatellites and insertion/deletion polymorphisms with $\delta > 0.3$ were called ‘Ethnic-Difference Markers’ (EDMs) (Collins-Schramm et al. 2002). Additionally, markers with high $\delta$ and very high log likelihood allelic ratio (LLAR: log of the ratio of allele frequencies for alleles at a locus in two populations) between populations have been designated ‘Population Specific Alleles’ (PSAs) (Shriver et al. 1997). This report followed from earlier work where markers with large allele frequency difference were identified to be appropriate for admixture studies (Chakraborty et al. 1991; Stephens et al. 1994) and most (0.95) of the arbitrarily identified biallelic markers had $\delta > 0.5$ (Dean et al. 1994). Thus, the authors proposed that PSAs should have $\delta > 0.5$ and also indicated that for multiallelic (no. of alleles > 2, e.g. microsatellites) loci, a composite $\delta$ could be estimated as one half the summation of the absolute value of allelic frequency differences for all alleles at that locus (Shriver et al. 1997). It has also been shown that markers with lower $\delta$ values, ($\delta \sim 0.3$), can provide up to 80 % power for detecting associations at distances of 5 cM with a large enough sample size ($N = 1000$) (Pfaff et al. 2001). These authors also suggested referring to markers with large allele frequency differences as ‘Ancestry Informative Markers’ (AIMs). Given that the central feature of these markers is the ancestry information content the term AIM more accurately describes these markers and does so using language that is less likely to be misunderstood and
misinterpreted (Molokhia et al. 2003; Rosenberg et al. 2003; Shriver et al. 2003; Pfaff et al. 2004).

**Marker information content**

It is well established that only 5–15% of the total genetic variation results from differences among human populations (Nei 1987; Cavalli-Sforza et al. 1994; Deka et al. 1995). Moreover, most alleles are shared between populations, and alleles common in one population are also common in other populations. Thus, most genetic markers are unaffected by admixture and it is imperative to choose markers that show high levels of $\delta$ between the parental populations. Theoretically, any marker that has large allele frequency difference between ancestral populations can be used for estimating individual ancestry and Admixture Mapping (AM). These can include SNPs, microsatellites, insertion/deletion polymorphisms (DIPs). Apart from $\delta$ other measures have been proposed to measure marker information content. These are summarized in Table 1.4 and briefly discussed here. Ancestry information can be measured by score variance ‘$f$’ (Fisher information) (McKeigue 1998; Molokhia et al. 2003; Shriver et al. 2003; Pfaff et al. 2004) by the expected log likelihood ratio (Kullback-Leibler information) (Rosenberg et al. 2003; Smith et al. 2004) which is contributed by typing a gamete at the marker locus. All these measures can be calculated from ancestry-specific allele frequencies and are expressed as the proportion of information that would be extracted by a perfectly informative marker, against a uniform prior distribution. The symbols “$f$” (McKeigue 1998) and $I_n$ (Rosenberg et al. 2003) have been used to designate Fisher information and Kullback-Leibler information respectively. Both methods rank markers similarly for ancestry information content although the absolute values for such measures differ (McKeigue 2005; McKeigue personal communication). In this thesis I have used $\delta$ values along with Fisher information to express marker informativeness. Fisher information content “$f$” is related to locus-specific $F_{ST}$ distance and is a value representative of the differentiation between two populations at a single locus. Specifically,
and is equivalent to Wahlund’s standardized variance for allele frequency. For example, if the average $F_{ST}$ distance between Europeans and West Africans is $\sim 0.15$ (Cavalli-Sforza et al. 1994) then the average $f$ value between these two populations is $\sim 0.08$.

Simulation studies for estimating the information content of markers with varying levels of $f$ have shown that for 1,000 markers with average information content for ancestry at 0.4 between two ancestral subpopulations, approximately 80% of the information about ancestry can be extracted from an initial genome screen. (McKeigue 1998; McKeigue et al. 2000; Hoggart et al. 2003; McKeigue 2005). After initial identification of regions showing admixture, more markers can be typed in these regions to increase extraction of information to nearly 100% (McKeigue 1998; Hoggart et al. 2004; McKeigue 2005).
Table 1.4: Measures of Marker ancestry information content

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
<th>Formula</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>δ : Measure of absolute allele frequency difference</td>
<td></td>
<td>$</td>
<td>p_1 - q_1</td>
</tr>
<tr>
<td>$F_{ST}$</td>
<td>Correlation between homologous genes compared to a random pair of genes in a population.</td>
<td></td>
<td>(Excoffier 2001)</td>
</tr>
<tr>
<td>$f$ (Fisher information content)</td>
<td>$f = F_{ST}/(2-F_{ST})$</td>
<td></td>
<td>(McKeigue 1998; Molokhia et al. 2003)</td>
</tr>
<tr>
<td>$I_a$ (informativeness for assignment)</td>
<td>Expected log of the ratio of the likelihood that an allele is assigned to one population to the likelihood of the allele being assigned to the average population (average of k subpopulations).</td>
<td></td>
<td>(Rosenberg et al. 2003)</td>
</tr>
<tr>
<td>$I_m$ (Fisherian information)</td>
<td>Derived by taking into account δ, allele frequencies in ancestral populations and admixture proportions. Estimated as the inverse of variance of maximum likelihood estimator of admixture proportions.</td>
<td></td>
<td>(Pfaff et al. 2004)</td>
</tr>
</tbody>
</table>

Recently it was suggested that with the availability of multiplexed assays and reduced genotyping costs it will be easy to score tens or hundreds of thousands of SNPs and it may become unnecessary to pre select SNPs for ancestry informativeness (Montana and Pritchard 2004). Simulation studies by Montana and Pritchard (2004) showed that the density of markers required to achieve a certain level of information content is inversely proportional to the ancestry-information content of the individual markers. McKeigue (2005) argues that by following Montana and Pritchard’s (2004) suggestion and using an unselected marker panel in a population with mixed European and West African ancestry one would achieve an average $f$ value of $-0.08$ with 10,000 unselected markers. However, by using a panel of only 2,000 ancestry informative markers whose average $f$ value is 0.4 one would get equivalent information. Thus
McKeigue (2005) argues for preselecting ancestry informative markers. Additional benefits to selecting informative markers include obvious lower genotyping costs, exclusion of markers whose allele frequencies vary within continental populations, adequate marker spacing to exclude allelic association within subpopulations and reduced of computational burdens.

**Available marker panels:**

Recent studies by several groups have focused on identifying panels of markers suitable for admixture studies as well as in developing measures of ancestry information content of markers. One study screened 744 microsatellite markers for composite $\delta$ values and LLAR in four different populations and identified a genome spanning set of 315 markers (average spacing 10 cM, $\delta>0.3$) for mapping in African Americans and 214 markers (average spacing of 16 cM, $\delta>0.25$) for mapping in Hispanics (Smith et al., 2001). A DNA pooling method was used to identify 151 AIMs (microsatellites and short insertion/deletion polymorphisms), with $\delta>0.3$ for mapping in Mexican-American populations and to distinguish between European-American and Indigenous-American contributions (Collins-Schramm et al. 2002). Ninety-seven AIMs were identified by the authors for mapping in African-American populations that have the added benefit of showing limited variation within a selection of West African populations (Collins-Schramm et al. 2002b). Several SNP AIM panels have been reported over the past few years (Shriver et al. 1997; Parra et al. 1998; Akey et al. 2002; Shriver et al. 2003; Bonilla et al. 2004b; Bonilla et al. 2004a). Recently ~ 3000 SNP AIMs were reported for studying African-American populations (Smith et al. 2004) after screening 450,000 SNPs for which allele frequencies were available. Allele frequencies for these AIMs in European, West African, Indigenous American and East Asian populations were presented for a subset of these. Additional resources are available for obtaining marker frequency, and genotype and haplotype information, from The SNP Consortium (TSC; http://snp.cshl.org), the National Center for Biotechnology Information’s ‘dbSNP’
website (http://www.ncbi.nlm.nih.gov/SNP), the Marshfield Database (http://research.marshfieldclinic.org/genetics/ Default.htm) and the HapMap project (www.hapmap.org) where over 1,000,000 allele frequencies are reported in 30 family trios from US who are primarily of northern and western European descent, 30 family trios from Nigeria, 45 unrelated individuals from Japan and 45 unrelated individuals from China.

While numerous markers are now available for admixture studies, some caution needs to be exercised when selecting suitable markers for any particular study. Most of the marker panels proposed to date have focused on African Americans as the admixed population of choice (Parra et al. 1998; Smith et al. 2001; Collins-Schramm et al. 2002; Smith et al. 2004), while a few have also focused on Hispanic populations (Bonilla et al. 2004b; Collins-Schramm et al. 2004; Bonilla et al. 2004a). However, these panels are not exhaustive and there is ample scope and need for developing new panels. Second, these panels have been developed assuming certain demographic characteristics. For instance, panels for African Americans typically have markers that distinguish between European and West African populations. While such a panel is likely to be sufficient for most African-American populations, it is not unreasonable to assume that depending on geographical location and demographic history, some African-American populations may also have substantial Indigenous-American ancestry. In such cases it may be important to include markers for measuring Indigenous American ancestry as well as those measuring European and West African ancestry. Again, populations of West African ancestry in other parts of the world, like African Caribbeans in the UK may have demographic histories different from African Americans in the US. The problem intensifies further for populations like the Trinidadians where there is evidence of five-way admixture between Europeans, Africans, Chinese, Indians (from India) and Indigenous Americans. Using the available marker panels may be an appropriate place to start the analysis, but it is obvious that further efforts should be made to identify markers that are informative for each ancestral population. A related issue when more than two ancestral populations are assumed to have contributed to the admixed population being studied relates to the choice of marker panels. Thus far no examination has been made whether it is necessary to
identify markers that distinguish between all populations equally. For instance, suppose a Latino population is being studied. In this case, it may be important to select markers such that the markers taken together provide equal power to distinguish between each pair of ancestral populations, European, African and Indigenous American.

A reasonable hypothesis seems to be that a balanced marker panel will provide more reliable estimates, since there will be less bias along any one ancestry axis brought about by choice of marker panels. For example, when analyzing a Latino population which has substantial European, West African and Indigenous American ancestry, using a marker panel which has higher distinguishing power between African and non-African populations compared to distinguishing between different non-African populations (European and Indigenous American in this case), estimates of non-African ancestry may be biased along either the European or Indigenous American ancestry axis.

**Ancestry component–phenotype correlations**

Traits and diseases more prevalent in one population than in others are amenable to admixture analysis and some examples are listed in Table 1.1. Most of the diseases shown in this table have a complex etiology affected by multiple genes and environmental factors. Earlier studies (Gardner et al. 1984; Long et al. 1991b) focused on admixed populations as units of analysis in exploring relationships between ancestry and phenotypes (Chakraborty and Weiss, 1988) These authors showed that non-insulin-dependent (Type 2) diabetes mellitus (NIDDM) prevalence is correlated with admixture proportions among a selection of populations with varying levels of Indigenous American ancestry. Data like these provide compelling evidence for the existence of frequency differences in risk modifying alleles, but such data have not been collected for many diseases. Another related approach is to test for individual admixture–phenotype correlations within an admixed population. Correlations between ancestry and phenotypes have been reported in a number of studies (Gardner et al. 1984; Hanis et al. 1986; Long et al. 1991b; Fernandez et al. 2003; Gower et al. 2003; Molokhia et al. 2003;
Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004a). A prerequisite for testing ancestry/phenotype correlations is the presence of stratification related to admixture, which will be evident in variation in individual ancestry levels. Figure 1.5 shows the distribution of BGA estimates from a sample of Puerto Rican women from New York city (Bonilla et al. 2004a), Figure 1.6 shows BGA distribution in a sample of Hispanics from the San Luis Valley, Colorado (Bonilla et al. 2004b). Figure 1.7 shows distribution of BGA in a sample of Mexicans from Tlapa, Mexico. Substantial variation is observed in all three samples. With the San Luis Valley sample, more variability is observed on the European–Indigenous American axis, while the New York sample is more variable on the European–West African axis. Following the argument of Chakraborty and Weiss (1986), admixture proportions should be correlated with diseases/traits that differ in populations due to underlying genetic differences. In each of these population samples, strong positive correlation was observed between individual ancestry and skin pigmentation measured as melanin index ‘M’ or lightness index ‘L’ (Figures 1.5, 1.6 and 1.7). A significant negative correlation was also observed between the proportion of West African ancestry and bone mineral density (BMD) in the Puerto Rican sample (Bonilla et al. 2004a). West African ancestry proportion and skin pigmentation (measured as melanin index) are also correlated in African Americans from Washington DC and African Caribbeans from the UK, but not in European Americans from State College, Pennsylvania (Shriver et al. 2003). Recently, correlations have been observed between proportion West African ancestry and lower insulin sensitivity, higher fasting insulin and acute insulin response to glucose in a combined sample of African-American and European-American children (Gower et al. 2003). In a separate sample of African-American females, West African ancestry is associated with body mass index, fat mass, fat-free mass and BMD (Fernandez et al. 2003). It is important to keep in mind that ancestry–phenotype correlations are dependent on both the existence of functional alleles at different frequencies in parental populations, and significant stratification related to admixture. This stratification is caused by the variation in individual ancestry proportions among individuals within a population. Although most admixed populations tested to date are structured, there is variation in the amount of stratification present, and this
structure should be tested for explicitly when investigating a new population. (Parra et al. 2001; Pfaff et al. 2001; Pfaff et al. 2002).

Figure 1.5: Correlation between % African ancestry and Melanin index in Puerto Ricans from New York City. Proportional ancestry was measured using 35 AIMs. (Source: Bonilla et al. 2004a)
Figure 1.6: Correlation between proportional ancestry and skin pigmentation in Hispanics from San Luis Valley; ancestry estimated using 22 AIMs. (Source: Bonilla et al. 2004b)
Methods for detecting stratification

Several methods have been developed to detect (McKeigue et al. 2000; Pritchard et al. 2000; Parra et al. 2001; Pfaff et al. 2001; Pfaff et al. 2002; Hoggart et al. 2003; Bonilla et al. 2004) and adjust (Devlin and Roeder 1999; Pritchard et al. 2000; Parra et al. 2001; Pfaff et al. 2001; Pfaff et al. 2002; Hoggart et al. 2003; Shriver et al. 2003) for stratification within a population sample. Some of these methods are briefly summarized here. 1) The Genomic Control method tests for the effects of population stratification on the candidate gene along with numerous other loci, which are not related to the disease (Devlin and Roeder 1999). The basic principle of this method is that if population stratification affects the candidate gene, it will also affect unrelated loci, which have no
effect on the disease, and variance inflation is constant across unrelated loci in the genome. Trend tests have been constructed for this purpose. In presence of population stratification, the variance of the trend test will be inflated. The variance inflation of the trend test for the null loci can be used to adjust that of the trend test for the candidate gene. 2) The Structured Association test (Pritchard et al. 2000) uses a set of unlinked genetic markers to infer details of population structure, namely deviation of Hardy Weinberg equilibrium and linkage equilibrium, and estimates the ancestry of sampled individuals to detect subpopulations in the study sample. This information is then used to test for associations within subpopulations. This is implemented in the programs STRUCTURE and STRAT. 3) The $D_e/D_0$ test examines the relationship between the observed LD and the predicted ALD between unlinked marker pairs for detecting structure within the sample (Pfaff 2001; Bonilla 2003). Simulation studies have shown that significant ALD between unlinked marker pairs exists when there has been continuous admixture from ancestral populations (Pfaff, 2001). In contrast, less than 5% of unlinked markers show an association after 15 generations since an admixture event followed by isolations of the hybrid population and ancestral populations. However, in its present form this test is applicable only when two ancestral populations have contributed to the admixed sample, such that initial levels of LD ($D_0$) can be easily estimated. 4) The Individual ancestry correlation test (IACT) examines the correlation in ancestry estimates obtained by non-syntenic marker panels using either random half sets of markers (Pfaff et al. 2001) or by taking all markers on the even chromosomes and all markers on the odd chromosomes (Shriver et al. 2005). Correlation in estimates obtained with the non-overlapping marker panels is a sensitive indicator of presence of stratification in the population. 5) Tests for unlinked marker pair correlations have been used to test for stratification. By chance only 5% of unlinked marker pairs are expected to show an association. If more than 5% of unlinked marker pairs show an association it is taken as evidence of stratification (Bonilla et al. 2004). A Bayesian method specifically developed for controlling admixture stratification while taking into account variation in individual admixture proportions and uncertainties associated with such estimates has been proposed (Hoggart et al. 2003).
Methods for Admixture Mapping:

Genetic analysis of phenotypes and diseases has traditionally followed one of two approaches: family-based linkage analysis and population-based association studies. While in linkage analysis it is the co-segregation of alleles in families that is measured, population-based studies measure non-random associations between phenotypes and alleles in population samples. Linkage analysis has proven to be immensely successful as a means of identifying genes for > 1,400 single gene diseases with simple Mendelian inheritance (e.g. see OMIM database). Complex diseases are multifactorial, polygenic and often characterized by late age of onset, incomplete penetrance, locus and phenotype heterogeneity and environmental exposures and, despite significant efforts, has only had moderate success in family-based mapping (for example the information from the Family Blood Pressure Project which was set up to use family based methods for mapping genes underlying hypertension).

Linkage Disequilibrium (LD) is an important aspect of genetic association studies and is generated in a population through mutation, selection, drift, non-random mating and admixture. Allelic associations due to LD are significant and are correlated with physical distance within small genomic regions but decay over time due to recombination (Jorde 1995; Huttley et al. 1999; Ardlie et al. 2002). Allelic associations can result either from direct functional effects of the alleles tested or indirectly through non-random associations between the allele measured and nearby functional alleles. Since functional alleles in most genes are still unknown and are indeed the object of research, LD is an important feature of how genes can be screened for alleles that alter disease risk. Thus, there has been substantial focus on the extent of LD across the genome and the definition of statistical methods for disease gene mapping using LD (Jorde 1995; Risch and Merikangas 1996; Pritchard and Przeworski 2001; Weiss and Clark 2002). LD-based association studies have been successful in both fine scale mapping (Kerem et al. 1989; MacDonald et al. 1999) and initial disease gene mapping in homogeneous populations that have undergone recent bottlenecks e.g. Hirschspring disease in Mennonites (Puffenberger 2003), Bardet–Beidl syndrome in Bedouins (Sheffield et al. 1994). In
large cosmopolitan populations, however, LD may be difficult to detect when the mutation is old, since the amount of remaining LD may be small. Additionally, false-positive associations due to population stratification are important confounders in any type of association study including LD-based association studies. LD due to admixture offers an opportunity to apply many of the general LD based methods in a unique situation. The characteristics and uses of ALD have been previously discussed in the introduction section. When studying ALD, issues related to population stratification are of major concern.

Non random mating with respect to individual ancestry proportions results in variation in individual admixture and population stratification, which in turn can inflate the number of significant associations that are observed (Ewens and Spielman 1995; Rybicki et al. 2002; Molokhia et al. 2003) and is a potential confounder in association studies (Lander and Schork 1994; McKeigue 1998; Thomas and Witte 2002; Hoggart et al. 2003). Using individual ancestry as a conditioning variable in analysis of variance tests, it is possible to eliminate association of the trait with unlinked alleles (Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004a). The Bayesian approaches implemented by (McKeigue 1998; McKeigue et al. 2000) and (Pritchard et al. 2000) offer an advantage over classical maximum likelihood based methods (Chakraborty 1986; Chakraborty and Weiss 1986; Hanis et al. 1986) by allowing for missing genotype and ancestry data and modeling admixture hierarchically. Methods have also been developed to control for parental admixture (McKeigue 1998) and to account for uncertain BGA estimation (McKeigue et al. 2000; Hoggart et al. 2003).

Theoretical and experimental studies have explored the parameters that characterize and affect admixture studies. (Briscoe et al. 1994; Dean et al. 1994; Stephens et al. 1994; Parra et al. 1998; Lautenberger et al. 2000; Parra et al. 2001; Pfaff et al. 2001). Early researchers viewed admixture approach as a simple extension of LD mapping methods and the acronym MALD (Mapping by Admixture Linkage Disequilibrium) was proposed (Briscoe et al. 1994; Stephens et al. 1994) to designate the mapping method proposed originally by Chakraborty and Weiss, which exploited the long range allelic associations created through ALD. Parameters critical for MALD
include the genetic distance between ancestry markers and the disease locus ($\theta$), number of generations since admixture ($t$), proportion of admixture ($m$) from one parental population, the allele frequency differential ($\delta$) between parental populations, and sample size ($N$) (Chakraborty and Weiss 1988; Briscoe et al. 1994; McKeigue 1998). Simulation studies suggest that sample sizes of 200–300 patients, typed for 200–300 evenly spaced markers, each having allele frequency differentials 0.3, have a 95 % chance of locating the causative gene, when there has been no new admixture from the parental population in the last four generations and no other sources of population structure or sample heterogeneity (Briscoe et al. 1994; Stephens et al. 1994).

Other approaches proposed for using admixture include a method based on the transmission disequilibrium test (TDT) (Ewens and Spielman 1995) that assesses excess transmission of alleles derived from high-risk ancestors to affected offspring of parents who are heterozygous at the marker locus, containing one allele from each of two ancestral populations (McKeigue 1997). A second TDT-based likelihood approach was developed that compared the transmission of haplotypes with non-transmission in affected offspring in an admixed population following a multipoint method. A likelihood statistic is calculated to determine the significance of various models under different scenarios (Zheng and Elston 1999).

One fundamental limitation of MALD as initially described and in its early extensions, is the effect of stratification on causing false-positive association (Chakraborty and Weiss 1988; Briscoe et al. 1994; Stephens et al. 1994). The TDT is one means of correcting for this stratification. Another is by conditioning on parental admixture (McKeigue 1998; McKeigue et al. 2000). This method provides an excellent example for the dependence between the descriptions of ancestry that has been discussed previously. For mapping purposes the description of interest is genomic ancestry in an individual, but in this approach, the inference is based on genealogical ancestry for that person. This further highlights the difficulties in ascertaining genomic ancestry for an individual and illustrates that certain assumptions need to be made in order to make any meaningful inferences. In this method information about ancestry from marker data at all loci are combined to estimate ancestry of alleles at each locus. When allelic ancestry at
marker loci is known, this approach is analogous to a linkage analysis; hence the term Admixture Mapping (AM) is more appropriate than MALD for describing this method and to distinguish it from LD approaches. (McKeigue 1998; McKeigue et al. 2000; Hoggart et al. 2003; Hoggart et al. 2004). The underlying variation in ancestry of chromosomes of mixed descent is modeled to extract all of the information about linkage that is generated by admixture. For example, where a locus is assumed to account for variation in skin pigmentation between two parental groups, e.g. West Africans and Europeans, individuals can be classified according to whether they have 0, 1 or 2 alleles of West African descent at this locus. By comparing these three groups for mean pigmentation level, holding all other factors constant, variation in pigmentation can be observed depending upon the number of alleles of West African ancestry in an individual. Controlling for parental admixture eliminates association of the trait with ancestry at unlinked loci. By removing the background effects of ancestry, it is possible to observe the locus specific effects on a trait/disease (Shriver et al. 2003; Bonilla et al. 2004a).

Allelic ancestry at a locus is inferred from the marker by using the conditional probability of each allelic state given the ancestry-specific allele frequencies. A complex hierarchical model with many nuisance parameters is used to model the distribution of admixture in the population. This is implemented using the ADMIXMAP program (available at http://www.ucd.ie/genepl), which follows a Bayesian approach with Markov chain simulation, and incorporates the admixture of each individual’s parents and the random variation of ancestry on chromosomes inherited from each of the parents in the model (McKeigue et al. 2000; Hoggart et al. 2003; Hoggart et al. 2004).

Some other groups have also recently proposed Bayesian frameworks for multipoint statistical modeling of genotype data from admixed individuals. In one of these methods “ancestry association” is based simply on the deviation of locus ancestry in a sample from the genome average (Montana and Pritchard 2004). This method is similar to that proposed earlier (Zhu et al. 2004), which calculates a Z statistic to infer deviations from mean genomic ancestry at a locus. The program MALDSOFT has been developed to implement this approach (Montana and Pritchard 2004). The main different between the approaches developed by Zhu et al., 2004 and that proposed by Montana and
Pritchard, 2004 appears to be the method by which admixture in the sample is estimated. MALDOSOFT has been developed to work in conjunction with the software STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) and make locus specific inferences based on the results from this program. Whereas, the approach proposed by Zhu et al. (2004), does not depend on a pre-existing program, instead it can use estimates of admixture proportions obtained by any method to infer variation in admixture proportions at a specific locus. A second method was proposed by (Patterson et al. 2004) uses a Hidden Markov Model to estimate ancestral origins of loci when several linked loci are available to create composite loci. The program ANCESTRYMAP (Patterson et al. 2004) has been developed to implement this approach.

**Specific Aims of this project:**

The **first specific aim** is to estimate individual ancestry proportions in several population samples using autosomal markers and to detect and characterize admixture stratification in these populations using multilocus genotype data. Examination of the distribution of ancestry proportions can provide important insights into the population structure. Three different methods have been used for this purpose: a Maximum Likelihood method (Hanis et al. 1986) and two Bayesian methods implemented in the programs STRUCTURE (Pritchard et al. 2000) and ADMIXMAP (McKeigue et al. 2000). Several tests for detecting admixture stratification including the IACT, unlinked marker-pair correlation test, $D_i/D_0$ test, and test for subpopulations implemented within STRUCTURE and ADMIXMAP programs are used for this purpose.

The **second specific aim** is to test for ancestry-phenotype correlations in samples of admixed individuals. One way to decompose and distinguish between environmental and genetic determinants of disease risk is to identify an association between admixture proportions and phenotype of interest. Detecting such an admixture-phenotype association provides support for a genetic basis underlying the phenotype.
In addition, where applicable, locus specific associations and ancestry-specific association of phenotypes is tested.

The **third specific aim** of this project is to compare two different methods mentioned in specific aim one and that are commonly used for estimating individual admixture, using computer simulations. Under each method, several of the parameters specific to the method are tested. Specific questions I have investigated are: 1) Does a panel of markers where information content for all ancestry components is balanced (i.e. nearly equal) provide more reliable estimates compared to an unbalanced panel? 2) Do Bayesian methods provide more reliable estimates than ML when the marker panel is balanced? 3) What is the effect of including ancestral populations in the Bayesian analyses vs. not including any ancestral populations? 4) When using STRUCTURE, what if any is the effect of specifying the “POPFLAG” option when individuals representing ancestral populations are included and identified as the ancestral populations in the analyses? 5) When using STRUCTURE what if any is the effect of using a model with separate alpha parameter vs. same alpha parameter for the ancestral population?

The overall aim of these studies is to investigate whether reasonable assumptions can be made under certain conditions and to try to identify if one method provides more reliable estimates of individual ancestry than others under certain situations.

Several data sets became available during the course of this project and were used for these studies. Most studies on individual admixture to date have been carried out with small sets of markers (Parra et al. 1998; Williams et al. 2000; Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004a; Reiner et al. 2005). The data sets were available through existing collaborations between Dr. Shriver’s laboratory and the laboratories of Dr. N. Schork (University of California, San Diego), Dr. B. Falkner (University of Pennsylvania), Dr. J. Berenson (Louisiana State University) and Dr. J. Gelernter (Yale University).

In chapter 2 individual admixture proportions in several US populations are described. These include one European-American sample, one African-American sample two different Latino samples from different geographic locations. These population
samples were genotyped for a panel of 36 microsatellite and 1 SNP marker. Admixture stratification was detected in the African-American and Latino samples. In addition, substantial geographic heterogeneity was detected between the two Latino samples.

Chapter 3 describes admixture structure in two African-American samples, one from Philadelphia, PA and one from Bogalusa, LA. These samples were genotyped for smaller SNP AIM panels. These AIMs have been described previously (Parra et al. 1998; Williams et al. 2000; Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004a) and are part of the core set of AIMs developed in Dr. Shriver’s laboratory. Estimates of individual admixture were calculated in each case and are also described. Several phenotypic measurements are available for these samples, including measures of blood pressure, blood insulin and glucose, levels of HDL, LDL and triglycerides, BMI and anthropometric measures. The combined sample size in this study was 376 unrelated individuals. Only moderate effects of overall ancestry on phenotypes were detected. Availability of well-characterized phenotypes along with a panel of well-characterized SNP AIMs allowed testing for ancestry-phenotype correlations. Admixture mapping of BMI, blood pressure, Blood glucose and insulin, HDL, LDL and triglyceride levels was done using two different approaches. First, the ANOVA/IAE (analysis of variance after controlling for individual admixture) method was used for detecting linkage to genes after controlling for admixture stratification. This method has been used previously for mapping genes that influence variation in normal skin pigmentation (Shriver et al., 2003). Secondly, Bayesian methods implemented in the ADMIXMAP software were used for mapping. Strong allelic association and ancestry association was observed for the locus SCG30050 (rs3317) with BMI in the combined sample.

Chapter 4 describes multilocus microsatellite genotype data set on ~13,000 individuals from several European-American and African-American populations and one Mexican-American population was used to explore variation in individual admixture proportions in several large population samples. A common panel of 312 microsatellites was typed in these individuals that were collected as part of the Family Blood Pressure Program. Tests for detecting admixture structure in populations were conducted and these are also described in this chapter. Evidence of admixture stratification is seen in all the
African-American and Mexican-American samples but not in the European-American samples. Several phenotype measures including anthropometric variables, blood glucose and insulin, blood HDLc, LDL and triglyceride levels and BMI were also available for a subset of these individuals. Ancestry-phenotype correlations have been tested and are described in this chapter.

A second data set consisted of genotypes on 11,077 single nucleotide polymorphisms on several populations including a sample of 42 African Americans, 41 European Americans and 20 Puerto Ricans. Estimates of individual admixture and tests for admixture stratification were conducted in these samples. The results and analyses are described in Appendix D and have been published as part of a larger report describing genetic variation in several populations (Shriver et al. 2005).

Several studies have previously used one or more of the methods mentioned in Specific Aim 1 for inferring individual admixture proportions in empirical studies of human populations (Williams et al. 2000; Parra et al. 2001; Shriver et al. 2003; Bonilla et al. 2004a; Reiner et al. 2005; Zhu et al. 2005). During initial phases of this project, I have used two different methods for inferring individual ancestry. While correlations in ancestry estimates obtained with different methods are high, there are also some interesting differences. However, no studies have been done to date using computer simulations where samples of individuals of known ancestry levels have been used to compare the reliability of the estimates obtained by these methods through. Chapter 5 of this thesis describes a simulation study that has been performed to test the reliability of admixture estimates obtained with different methods. While simulating a population sample may appear relatively simple, developing a model which realistically depicts a population sample was challenging. An algorithm was developed by which several of the parameters that may affect individual admixture can be controlled and/or varied. Simulations were then performed to test for the reliability of the estimates obtained.

Chapter 6 is a comprehensive chapter describing overall conclusions from these studies and directions for future analyses.

In addition 5 appendices are included. Allele frequencies for microsatellites described in Chapter 2 are included in Appendix A. The simulation program written for
simulating individuals based on proportional genealogical ancestry is (Sim_Sample.pl) is provided in Appendix B. A second simulation scheme where alleles were tagged and followed through several generations of admixture (Sim_Sample_tag.pl) is presented in Appendix C. This simulation program has been used in Chapters 4 and 5. Analysis of individual ancestry in three different populations; using data from the Affymetrix 10K SNP Chip is presented in Appendix D. The code for calculating maximum likelihood estimates is provided in Appendix E.
References


Williams RC, Long JC, Hanson RL, Sievers ML, Knowler WC (2000) Individual estimates of European genetic admixture associated with lower body-mass index,
Zoratti R (1998) A review on ethnic differences in plasma triglycerides and high-density-lipoprotein cholesterol: is the lipid pattern the key factor for the low coronary heart disease rate in people of African origin? Eur J Epidemiol 14:9-21
Chapter 2

Measurement of admixture proportions and description of admixture structure in different US populations

Introduction:

Understanding the genetic structure of populations is of genetic and anthropological interest. It is also of practical importance for the valid control of population stratification for case-control association studies. Each individual in a population has a genetic history that is likely to differ from that of other individuals in that population. Thus, evaluating genetic variation within and between populations has important consequences for both forensic and medical genetics. At the time of this writing the US has three predominant population groups: Individuals who identify themselves as “White”, “Black or African American” and “Hispanic or Latino” (US Census Bureau, 2000). According to recent estimates African Americans represent 12.3% of the total US population. This categorization is used to designate a range of people of exclusively African origin who have little or no mixed ancestry, as well as people of African descent who have substantial non-African ancestry. Latino (or Hispanic) individuals represent 13.7% of total US residents and include Mexican Americans, Puerto Ricans, Cubans and other populations who speak Spanish or Portuguese; that is they are categorized based on language and cultural affiliations. These populations may, depending on the particular group sampled, have greater proportional Indigenous American (IA) and European (EU) and lesser West African (WAF) ancestry, while some have very high WAF ancestry as well. Other populations, including different Asian groups, are represented among US residents in smaller numbers.
Many diseases have prevalence differences among populations for which there is evidence that some of the difference is a reflection of group genetic differences -- for example, hypertension and prostate cancer in African Americans (Douglas et al. 1996; Hoffman et al. 2001; Hoffman et al. 2003) and diabetes in American Indians (Martinez 1993; Songer and Zimmet 1995) (For a more detailed summary see Table 1.1 of Chapter 1 of this thesis and Table 1 from (Halder and Shriver 2003). Understanding the genetic structure of populations and exploring the geographic heterogeneity may aid in developing a better understanding of the genetic processes and, eventually, the disease etiology, and help decompose the sources of variation (environmental, genetic and gene-environment interactions) that contribute to the disease risk. Understanding and characterizing such variation will also help the forensic community and efforts can be made to construct regional and phenotypically qualified databases.

The process of intermixture between populations (or individuals belonging to different populations) that have been geographically separated, results in a newly admixed population. Different models have been proposed to explain and model how the process of admixture occurs in populations (Long et al. 1991; Pfaff et al. 2001). The two extreme cases are “Hybrid Isolation” where admixture occurs in a single generation between different ancestral populations, and “Continuous Gene Flow” where admixture occurs through successive generations between one or more of the ancestral populations and the new admixed population. Contemporary admixed populations usually have histories in which both processes have contributed, thereby creating a compound distribution of individual ancestry proportions in the population. Investigating and understanding variation in ancestry levels within a population is important for many reasons, including being able to model population stratification. Stratification has been implicated in causing false positives in genetic association studies (Heiman et al. 2004; Koller et al. 2004) and different methods have been proposed to detect (Pfaff et al. 2001; Hoggart et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004) and control (Pritchard et al. 2000; Devlin et al. 2001; Hoggart et al. 2003) such stratification. Furthermore, under certain circumstances the process of admixture can create extensive linkage disequilibrium (LD) over a broader range than is usually observed in non-admixed
populations where allele frequencies are different between the ancestral populations. This extended LD that occurs in admixed populations was first identified as a potentially powerful tool for mapping disease genes by (Chakraborty and Weiss 1988).

Latinos and African Americans are the two largest groups with high level of mixed ancestry among the resident US populations. Previous studies have reported on admixture proportions in different African American populations (Parra et al. 1998; Parra et al. 2001), Latino populations (Williams et al. 1992; Sans 2000; Bonilla et al. 2004a; Bonilla et al. 2004b), Icelanders (Helgason et al. 2000), Tibeto-Burmans (Wen et al. 2004) and Gypsies (Gresham et al. 2001). Studies have also investigated the relationship of admixture with phenotypes (Gardner et al. 1984; Hanis et al. 1986; Long et al. 1991; Molokhia et al. 2003; Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004b; Reiner et al. 2005). While earlier studies used populations as the units of analysis to test for association between ancestry and phenotypes (Gardner et al. 1984; Williams et al. 1992), recent studies have focused on variation of individual ancestry proportions with phenotypes (Fernandez et al. 2003; Shriver et al. 2003; Bonilla et al. 2004a; Bonilla et al. 2004b; Reiner et al. 2005). A recurring theme in these studies has been the variability of group or population ancestry estimates in different populations, possibly reflecting their different demographic histories. While estimates of population admixture provide important information, examining the variation in individual ancestry estimates within the population will further enhance understanding of genetic structure of the population; moreover, this knowledge is a necessary prerequisite for structured association (Pritchard et al. 2000; Pritchard and Donnelly 2001) and related methods. Within different African American populations average proportional European (EU) ancestry ranges from $0.035 \pm 0.009$ in Gullah Sea Islanders of South Carolina to $0.22 \pm 0.016$ in New Orleans (Parra et al., 2001). Similarly, estimates of average ancestry have been reported for diverse Latino populations; for instance Mexican Americans from Arizona show $0.68 \pm 0.05$ EU, $0.29 \pm 0.04$ Indigenous American (IA) and $0.03 \pm 0.02$ West African (WAF) ancestry (Williams et al., 1992), Puerto Ricans from New York city show $0.53 \pm 0.03$ EU, $0.18 \pm 0.03$ IA and $0.3 \pm 0.02$ WAF ancestry (Bonilla et al., 2004a), Mexican-Americans from California show $0.6$ EU and $0.4$ IA ancestry (Henis et al. 1986). Moreover within any one
population there is a continuum of genetic admixture proportions. Many of these differences have not been tested for significance; it is possible that across populations, the differences observed are insignificant. In a recent study of ancestry proportions in different Latino populations in the continental United States, using six autosomal STR markers, the authors observed that Latino populations from the Eastern US had lower IA admixture compared to Latino populations in the Western US, while WAF ancestry was estimated to be similar across both regional groups (Bertoni et al., 2003). In a separate study using 9 STR haplotypes on the Y chromosome in different US populations of mixed ancestry, no evidence of geographic heterogeneity was observed among different European American, African American and Latino populations living in the US (Pfaff et al. 2001; Hoggart et al. 2003; Kayser et al. 2003). However, it is unlikely that a small panel of unselected markers will provide sufficient power to resolve ancestry proportions accurately. To investigate the genetic variation within and among populations, we have estimated proportional individual ancestry levels in four different US populations using a panel of 37 STRs and a single SNP locus (FY).

The populations included in this study comprise of four US resident populations, European Americans from Connecticut (henceforth referred to as EACT), African Americans from Connecticut (henceforth referred to as AACT) and two Latino populations, one from Connecticut, (referred henceforth as LCT) and one from California (henceforth referred to as LCA). Examination of variation in individual ancestry proportions is likely to provide important insights into the genetic structure in these populations. While certain assumptions are inherent in such estimates, the estimates nevertheless provide an initial basis for understanding the genetic variability in population samples. Variation in individual ancestry levels creates genetic substructure that if undetected and not controlled for, will lead to false positives and false negatives in genetic association studies (Pfaff et al. 2001; Hoggart et al. 2003). One way to control for spurious associations is by using estimates of individual ancestry proportions as a covariate in genetic association studies (Shriver et al. 2005). Previous studies on individual admixture in African Americans (Parra et al. 1998; Parra et al. 2001) and Latinos (Bonilla et al. 2004a; Bonilla et al. 2004b) have shown that there is variation in
individual ancestry proportions and associated stratification. Analysis of the AACT, LCT, and LCA samples will provide further information regarding the genetic structure in these populations. Ancestry proportions in European Americans are generally assumed to be more homogeneous and very few studies have been done to investigate variation in individual ancestry levels in European Americans (Shriver et al. 2003). The present data provides an opportunity to examine the genetic structure in a European American population sample using the same panel of markers as the African American and Latino samples.

In a recent study it was reported that a particular set of ancestry informative markers (AIMs) that distinguish between Europeans and Asians are equally informative for European-Indigenous American distinction (Collins-Schramm et al. 2004), and it was suggested that AIMs that distinguish between Europeans and Asians could be screened to select additional AIMs for admixture analysis in Mexican-American populations. We tested for this hypothesis using the current marker panel by genotyping the markers in a Thai population and comparing European-Asian $\delta$ values to European-Indigenous American $\delta$ values.

**Materials and Methods:**

**Populations:**

We studied 652 unrelated European American (EACT) and 228 unrelated African American (AACT) individuals from Connecticut (described previously by Yang et al. 2005), 102 Latino individuals from CT (LCT), and 89 Latino individuals from Southern California (LCA). Ancestral allele frequencies and genotypes were ascertained from a panel of 59 Spanish (representative European ancestral population, henceforth referred to as EU), 78 Mende (representative West African ancestral population, henceforth referred to as WAF), 88 Nahua from Mexico (representative Indigenous American ancestral
population, henceforth referred to as IA) and 154 Thai (representative Asian ancestral population, henceforth referred to as AS) individuals. The Nahua sample from Guerrero, Mexico was previously shown to have very low EU or WAF admixture (Bonilla, 2003) and hence can be used as an appropriate IA sample.

**Sample collection and genotyping:**

EACT, AACT and LCT samples were recruited at Yale University School of medicine and at the University Of Connecticut Health Center following protocols as described in Yang et al., (2005). LCA samples were recruited from Southern California following protocols as described in (Stein et al. 2004). In all cases individuals were self identified and verified by interviewer’s observations. Thai samples were recruited in Bangkok, Thailand. Nahua (IA) samples were collected from the city of Tlapa in Guerrero, Mexico and have been shown to have low EU or WAF admixture (Bonilla, 2003). The Mende (WAF) samples are from Sierra Leone and Spanish (EU) samples are from Valencia, Spain.

15 CODIS (Combined DNA Index System) markers that are routinely used for forensic analysis (D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S443, vWA, TPOX, D18S51, D5S818, FGA) and an additional 21 markers (D1S196, D1S2628, D2S162, D2S319, D5S407, D5S410, D6S1610, D7S640, D7S657, D8S272, D8S1827, D9S175, D10S197, D10S1786, D11S935, D12S352, D14S68, D15S1002, D16S3017, D17S799, and D22S274.) chosen to maximize the allele frequency difference between European and African populations and in some cases between IA and AS populations were chosen from the panel of markers described by (Smith et al. 2001). All markers were genotyped following the protocol described by Yang et al., (2005). In addition to the STRs, a single SNP, FY, was genotyped to further augment the difference between African and non-African populations. The 36 STR markers included in this panel have been optimized for genotyping in 2 lanes on the ABI PRISM 3100 semiautomated capillary fluorescence
sequencer. Data were scored using Genemapper (ABI). FY was genotyped by RFLP-PCR.

**Estimating individual ancestry:**

Individual ancestry was estimated using two different approaches a maximum likelihood (ML) method and a separate Bayesian method as implemented in the program STRUCTURE (Pritchard et al., 2000, Falush et al. 2003). In the ML method as previously described (Hanis et al., 1986, Chakraborty, 1986, Shriver et al., 2003, Bonilla et al., 2004a) the probability of observing a marker genotype, given ancestral allele frequencies at that locus, is computed. Summing over the logs of individual locus probabilities combines information across multiple loci. The admixture proportion that maximizes the probability of obtaining the observed genotype is the ML Estimate (MLE) of ancestry for the individual. Each genotype, at each locus, in each individual is treated as an independent instance. This method depends critically on the ancestral populations and ancestral allele frequencies used in the model. Perl scripts were written to calculate individual MLE of ancestry. Individual ancestry in EACT, AACT and the two Latino samples were calculated under a trihybrid model of admixture between EU, WAF and IA populations.

The program STRUCTURE uses a Bayesian Markov Chain Monte Carlo (MCMC) approach to model the admixture proportions of an individual. Posterior distributions of individual admixture proportions are generated given the observed marker genotypes, based on a hierarchical model of population admixture, individual admixture and locus ancestry. The algorithm jointly infers ancestry specific allele frequencies (the frequencies of each allele given the subpopulation of ancestry of the gene copy) and admixture proportions in individuals from the data. The models in STRUCTURE were set up to incorporate and identify ancestral populations by specifying the “Usepopinfo” option and allowing for admixture. Separate models were set up for analyzing each population, which included all three ancestral populations and one of the
study samples in a single iteration of the program. In addition allele frequencies were specified as being uncorrelated and a separate ‘alpha’ parameter (that characterizes the distribution of admixture proportions in each individual) was assumed for each population to allow for variation in ancestry proportions among individuals from different populations. The number of subpopulations (K) was specified from 1 to 3. For adequate convergence of the STRUCTURE algorithm, runs of different lengths were simulated. Final runs consisted of 40,000 burnin followed by additional 80,000 for parameter estimation. For all other options the default parameters were used.

Sample characterization:

Allele frequency estimates and exact tests for Hardy Weinberg equilibrium (HWE) were carried out using Genepop software (Rousset and Raymond 1995). All statistical analyses, besides those mentioned above were carried out using SPSS v10. Comparisons between group means were tested using two sample t-tests or one sample t-tests (when a single group was being tested for a specific mean value). Distribution of individual admixture between groups was compared using the Kolmogorov-Smirnov (KS) test for two samples. This is a non-parametric, empirical cumulative distribution test, which compares the hypothesis that two distributions are equal. Cumulative fraction plots of data were generated using a web based plotting tool available at www.physics.csbju.edu/stats/KS-test.html.

Marker information content for ancestry was estimated as composite δ (\(\delta_c\)) values which is an extension of biallelic δ, the difference in allele frequencies in two populations, to a multiallelic situation (Shriver et al. 1997). \(\delta_c\) is defined as:

\[
\delta_c = 1/2 \times \sum_{i=1}^{n} | f_{iA} - f_{iB} |
\]

where \(f_{iA}\) and \(f_{iB}\) are the frequencies of the \(i\)th allele in the two populations, A and B, being compared at a locus.
Detecting Admixture structure:

Two tests were used for detecting admixture structure in the sample. 1) The program STRUCTURE which tests for the presence of subpopulations within the sample being studied. For this, STRUCTURE was run by setting the predefined number of populations (K) for K=1, K= 2 and K=3 respectively. No ancestral populations were included in the sample and the model was set up allowing for admixture. Other parameters were kept the same as mentioned in the previous section. 2) The Individual Ancestry Correlation test was used in which the original marker panel is split randomly into two non-overlapping sets of markers (Pfaff et al. 2001). The procedure of randomly splitting markers is repeated 20 times and ancestry is estimated separately with each of the different (20 × 2) sets of markers. We then tested for correlation between estimates obtained with each half set of markers.

Simulation studies:

Simulated data sets were generated using the allele frequency estimates of the putative ancestral populations, i.e., the EU, WAF and IA. The general algorithm used for simulations is as follows: starting with a panel of allele frequencies from different population samples, individuals $X_1, X_2, X_3...X_i$ are generated from specified ancestral populations 1, 2, 3….i, “n” generations in the past, by randomly sampling alleles at each locus from the population “i” to generate genotypes for one individual $X_i$. All individuals generated in the first step of the program thus have 100% ancestry from any one of the ancestral populations. The numbers of individuals in the first generation who are from one ancestral population correspond to the expected ancestry proportion in the final simulated individual. For instance, to simulate an individual with an expected 0.75 EU ancestry and 0.25 WAF ancestry, four individuals are generated in the first step of the program; three are of 1.0 EU ancestry and one with 1.0 WAF ancestry. Individuals generated in the first step of the program are used as the putative grandparents of the
target individual with expected 0.75 EU ancestry. In the second step of the program, mating between individuals in the first generation is simulated, by iteratively choosing two individuals, without replacement, from this generation, and using alleles at each locus in these “parents” for generating locus specific genotypes of one “offspring”. At each locus random mating is simulated by choosing one allele from each parent. Thus, each parent has 50% chance of passing any one allele at a locus to an offspring. This allows for one individual to pass on 50% of alleles overall to the offspring. Since individuals in the first generation are used without replacement, two individuals are generated in the second generation. One of these individuals has both parents who are of EU ancestry, and this individual is expected to be 1.0 EU. The second individual in the second generation has one EU parent and one WAF parent and is thus expected to have 0.5 ancestry from each population. Once individuals in the second generation have all been generated, these serve as parents for the next generation. In the final step, mating is simulated between the individuals of the second generation following the same general scheme for choosing alleles from each parent, as described before. One allele at each locus is chosen from each “parent” and combined to generate the individual with expected 0.25 WAF and 0.75 EU ancestry. It is possible to control the number of generations of the simulation by changing the initial numbers of individuals generated in the first step of the program. So, for instance, to start a simulation that goes back 5 generations, we specify 32 ancestors and to start a simulation going back 10 generations we specify 1024 individuals in the first step of the program. Our algorithm also allows us to vary the number of generations so that we can simulate the same expected ancestral proportions but change the number of generations. For instance, in our first example above, we have simulated one individual with 0.75 EU and 0.25 WAF ancestry by starting with four grandparents, two generations in the past. Alternately, a second model can be specified by starting with 24 EU ancestors and 8 WAF ancestors five generations in the past to generate one individual with 0.75 expected EU ancestry and 0.25 expected WAF ancestry. A single iteration of the program can also be used to generate any “Y” numbers of independent individuals. The program generates each individual at a time, iteratively, starting from step one and repeating the entire process. For simulating a
sample of individuals with varying admixture proportions, separate models were set up, with different numbers of ancestral populations and for different numbers of generations in the past. Additional details for the simulation scheme are provided in Chapter 5 and the code for the perl script, sim_sample.pl that was written to perform these simulations is included in Appendix B. Simulated individuals belonging to any one population and with no admixture are expected to show 100 % ancestry from that population group. Departures between expected and observed values of ancestry estimates were examined by standard statistical tests.

Phylogenetic analysis:

Genetic distance between the populations was ascertained as pair wise F\textsubscript{ST} measures using Genepop software (Rousset and Raymond 1995). The F\textsubscript{ST} distance matrix was subsequently used in MEGA2 software (Kumar et al. 2001) to infer neighbor-joining tree of populations.

Results:

Allele frequencies and delta (δ\textsubscript{c}) levels:

The allele frequencies for all markers in all populations are shown in Appendix A. Allele frequencies in the EU, WAF, IA and AS populations were used in subsequent ML analyses for estimating proportional ancestry. All markers within each population were tested for consistency with HWE expectations. Markers that were not in HWE in each population are also indicated in Appendix A.

Marker information content for ancestry was measured as the allele frequency differences (δ\textsubscript{c}) between the ancestral populations and is shown in Table 2.1. As reported
previously (Yang et al. 2005), \( \delta_c \) and the index of informativeness for assignment, \( I_n \) (Rosenberg et al. 2003) have nearly identical information for this panel of markers. The total \( \delta_c \) value for pair-wise comparison was observed to be the highest between WAF and IA and lowest between the AS and EU. Markers with \( \delta \geq 0.3 \) have previously been shown as being suitable for ancestry estimation (Parra et al. 1998; Collins-Schramm et al. 2002; Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004). The proportion of markers that satisfy this criterion is 22/36 for EU-WAF comparison, 33/36 markers for WAF-IA comparison and 23/36 for the EU-IA comparison, 19/36 for AS-IA comparison, 29/36 for the AS-WAF comparison and 12/36 for AS-EU comparison. The marker duffy (FY), which is the only SNP in this panel, is an African ancestry-specific marker. This is the only marker to show nearly 100% allele frequency difference between the African and non-African ancestral populations. Positive correlations are observed between the \( \delta_c \) values for markers informative for EU-AS and EU–IA (\( R^2 = 0.329, P<0.01 \)) comparisons as well as WAF-AS and WAF-IA (\( R^2 = 0.5526, P < 0.005 \)) comparisons, supporting the results of a previous study (Collins-Schramm et al. 2004) which proposed that markers that distinguish between EU-AS may also have good distinction power between EU and IA and may be screened for additional markers that can be used for EU-IA distinction when studying Mexican American populations.
Table 2.1: Allele frequency differences ($\delta_c$) in different populations

<table>
<thead>
<tr>
<th>Markers</th>
<th>WAF-EU $\delta_c$</th>
<th>WAF-IA $\delta_c$</th>
<th>EU-IA $\delta_c$</th>
<th>AS-IA $\delta_c$</th>
<th>AS-WAF $\delta_c$</th>
<th>AS-EU $\delta_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSF1PO</td>
<td>0.232</td>
<td>0.211</td>
<td>0.140</td>
<td>0.084</td>
<td>0.173</td>
<td>0.113</td>
</tr>
<tr>
<td>D10S1786</td>
<td>0.470</td>
<td>0.545</td>
<td>0.516</td>
<td>0.423</td>
<td>0.321</td>
<td>0.477</td>
</tr>
<tr>
<td>D10S197</td>
<td>0.424</td>
<td>0.307</td>
<td>0.275</td>
<td>0.154</td>
<td>0.337</td>
<td>0.241</td>
</tr>
<tr>
<td>D11S935</td>
<td>0.756</td>
<td>0.583</td>
<td>0.337</td>
<td>0.367</td>
<td>0.590</td>
<td>0.335</td>
</tr>
<tr>
<td>D12S352</td>
<td>0.283</td>
<td>0.574</td>
<td>0.635</td>
<td>0.489</td>
<td>0.356</td>
<td>0.250</td>
</tr>
<tr>
<td>D13S317</td>
<td>0.265</td>
<td>0.519</td>
<td>0.452</td>
<td>0.376</td>
<td>0.565</td>
<td>0.300</td>
</tr>
<tr>
<td>D14S68</td>
<td>0.292</td>
<td>0.422</td>
<td>0.225</td>
<td>0.309</td>
<td>0.376</td>
<td>0.244</td>
</tr>
<tr>
<td>D15S1002</td>
<td>0.573</td>
<td>0.560</td>
<td>0.405</td>
<td>0.442</td>
<td>0.427</td>
<td>0.397</td>
</tr>
<tr>
<td>D16S3017</td>
<td>0.447</td>
<td>0.347</td>
<td>0.530</td>
<td>0.545</td>
<td>0.552</td>
<td>0.222</td>
</tr>
<tr>
<td>D16S539</td>
<td>0.181</td>
<td>0.231</td>
<td>0.246</td>
<td>0.232</td>
<td>0.155</td>
<td>0.231</td>
</tr>
<tr>
<td>D17S799</td>
<td>0.522</td>
<td>0.668</td>
<td>0.541</td>
<td>0.271</td>
<td>0.572</td>
<td>0.430</td>
</tr>
<tr>
<td>D18S51</td>
<td>0.320</td>
<td>0.310</td>
<td>0.204</td>
<td>0.235</td>
<td>0.339</td>
<td>0.271</td>
</tr>
<tr>
<td>D19S433</td>
<td>0.244</td>
<td>0.533</td>
<td>0.372</td>
<td>0.334</td>
<td>0.359</td>
<td>0.281</td>
</tr>
<tr>
<td>D1S196</td>
<td>0.270</td>
<td>0.337</td>
<td>0.160</td>
<td>0.243</td>
<td>0.409</td>
<td>0.272</td>
</tr>
<tr>
<td>D1S2628</td>
<td>0.624</td>
<td>0.479</td>
<td>0.339</td>
<td>0.196</td>
<td>0.642</td>
<td>0.357</td>
</tr>
<tr>
<td>D2S111</td>
<td>0.227</td>
<td>0.356</td>
<td>0.215</td>
<td>0.145</td>
<td>0.315</td>
<td>0.158</td>
</tr>
<tr>
<td>D2S274</td>
<td>0.339</td>
<td>0.377</td>
<td>0.372</td>
<td>0.400</td>
<td>0.494</td>
<td>0.367</td>
</tr>
<tr>
<td>D2S1338</td>
<td>0.447</td>
<td>0.352</td>
<td>0.424</td>
<td>0.292</td>
<td>0.262</td>
<td>0.335</td>
</tr>
<tr>
<td>D2S162</td>
<td>0.473</td>
<td>0.599</td>
<td>0.185</td>
<td>0.244</td>
<td>0.465</td>
<td>0.263</td>
</tr>
<tr>
<td>D3S1358</td>
<td>0.150</td>
<td>0.299</td>
<td>0.328</td>
<td>0.291</td>
<td>0.121</td>
<td>0.132</td>
</tr>
<tr>
<td>D5S407</td>
<td>0.369</td>
<td>0.419</td>
<td>0.371</td>
<td>0.316</td>
<td>0.387</td>
<td>0.242</td>
</tr>
<tr>
<td>D5S410</td>
<td>0.516</td>
<td>0.667</td>
<td>0.523</td>
<td>0.267</td>
<td>0.604</td>
<td>0.320</td>
</tr>
<tr>
<td>D5S818</td>
<td>0.207</td>
<td>0.510</td>
<td>0.331</td>
<td>0.332</td>
<td>0.350</td>
<td>0.227</td>
</tr>
<tr>
<td>D6S1610</td>
<td>0.456</td>
<td>0.517</td>
<td>0.412</td>
<td>0.470</td>
<td>0.390</td>
<td>0.297</td>
</tr>
<tr>
<td>D7S2469</td>
<td>0.360</td>
<td>0.359</td>
<td>0.359</td>
<td>0.563</td>
<td>0.535</td>
<td>0.437</td>
</tr>
<tr>
<td>D7S460</td>
<td>0.532</td>
<td>0.577</td>
<td>0.450</td>
<td>0.478</td>
<td>0.472</td>
<td>0.197</td>
</tr>
<tr>
<td>D7S657</td>
<td>0.682</td>
<td>0.503</td>
<td>0.393</td>
<td>0.260</td>
<td>0.268</td>
<td>0.429</td>
</tr>
<tr>
<td>D7S820</td>
<td>0.125</td>
<td>0.429</td>
<td>0.403</td>
<td>0.186</td>
<td>0.278</td>
<td>0.225</td>
</tr>
<tr>
<td>D8S1179</td>
<td>0.230</td>
<td>0.402</td>
<td>0.243</td>
<td>0.315</td>
<td>0.338</td>
<td>0.251</td>
</tr>
<tr>
<td>D8S1827</td>
<td>0.409</td>
<td>0.534</td>
<td>0.124</td>
<td>0.156</td>
<td>0.444</td>
<td>0.085</td>
</tr>
<tr>
<td>D8S272</td>
<td>0.458</td>
<td>0.552</td>
<td>0.428</td>
<td>0.409</td>
<td>0.362</td>
<td>0.269</td>
</tr>
<tr>
<td>D9S175</td>
<td>0.494</td>
<td>0.743</td>
<td>0.515</td>
<td>0.612</td>
<td>0.567</td>
<td>0.299</td>
</tr>
<tr>
<td>FGA</td>
<td>0.246</td>
<td>0.312</td>
<td>0.264</td>
<td>0.248</td>
<td>0.191</td>
<td>0.148</td>
</tr>
<tr>
<td>FY</td>
<td>0.968</td>
<td>0.980</td>
<td>0.012</td>
<td>0.537</td>
<td>0.986</td>
<td>0.018</td>
</tr>
<tr>
<td>TH01</td>
<td>0.421</td>
<td>0.483</td>
<td>0.369</td>
<td>0.206</td>
<td>0.280</td>
<td>0.460</td>
</tr>
<tr>
<td>TPOX</td>
<td>0.210</td>
<td>0.377</td>
<td>0.233</td>
<td>0.484</td>
<td>0.314</td>
<td>0.163</td>
</tr>
<tr>
<td>VWA</td>
<td>0.131</td>
<td>0.333</td>
<td>0.296</td>
<td>0.000</td>
<td>0.342</td>
<td>0.307</td>
</tr>
<tr>
<td>Total $\delta_c$</td>
<td>14.354</td>
<td>17.305</td>
<td>12.626</td>
<td>11.910</td>
<td>14.935</td>
<td>10.050</td>
</tr>
<tr>
<td>Average $\delta_c$</td>
<td>0.388</td>
<td>0.468</td>
<td>0.341</td>
<td>0.322</td>
<td>0.404</td>
<td>0.272</td>
</tr>
</tbody>
</table>
Estimates of individual ancestry proportions:

MLE of individual ancestry proportions in four US populations, the EACT, AACT and the two Latino samples are summarized in Table 2.2. Triangle plots illustrate the distribution of estimates of individual ancestry, where each point on the plot represents individuals with same genomic ancestry proportions. Each vertex represents one ancestral population and persons with 100% ancestry from a single population are expected to plot on the vertex corresponding to the ancestral group represented by that individual. The histogram accompanying the triangle plots further illustrate the distribution of admixture estimates in the samples.

<table>
<thead>
<tr>
<th>Population (N)</th>
<th>ML</th>
<th>STRUCTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IA</td>
<td>EU</td>
</tr>
<tr>
<td>EACT (652)</td>
<td>0.10 ± 0.1</td>
<td>0.84 ± 0.12</td>
</tr>
<tr>
<td>AACT (228)</td>
<td>0.08 ± 0.08</td>
<td>0.17 ± 0.14</td>
</tr>
<tr>
<td>LCT (102)</td>
<td>0.19 ± 0.16</td>
<td>0.59 ± 0.19</td>
</tr>
<tr>
<td>LCA (89)</td>
<td>0.36 ± 0.18</td>
<td>0.53 ± 0.2</td>
</tr>
</tbody>
</table>

Estimates of individual ancestry vary widely in this EACT sample as seen in Figure 2.1. Histogram shows the number of individuals with certain ancestry proportions. The histogram shows proportional WAF ancestry to be low for most individuals. ~ 52% of the individuals show >90% EU ancestry and these individuals all plot closer to the European vertex. 69% of the total population also show >0.05 IA ancestry and 57% showed >0.05 African ancestry. Further explanation and implications of this pattern of admixture are discussed later.
Figure 2.1: Distribution of Individual ancestry proportions in 652 European Americans from Connecticut (EACT) using Maximum Likelihood. WAF: West African, IA: Indigenous American; EU: European
In the AACT sample the average EU ancestry is 0.17 ± 0.14, which falls within the range for such estimates repeated previously (Parra et al. 1998). Distribution of proportional genomic ancestry levels for the AACT sample is shown in Figure 2.2. This sample also has small (0.08 ± 0.08) IA admixture. This proportion of IA ancestry is significant (P<0.0001) when tested against a null hypothesis of no IA admixture in the sample. Average WAF ancestry in the AACT sample is 0.75 ± 0.15. Substantial variation (variance in WAF ancestry is 0.39, in EU ancestry is 0.374 and in IA ancestry is 0.28) is seen in this sample, as shown in Figure 2.2, reflecting the recent history of admixture. The histogram shows that 8.3% of the sample has 1.0 WAF ancestry, 14.6% have 0 EU ancestry and 35% of the sample has 0 IA ancestry. Thus, in this AACT sample individual ancestry is distributed such that 50% have ancestry from at least one of the non-WAF ancestral populations and 42.1% have both EU and IA ancestry.
Figure 2.2: Distribution of Individual ancestry proportions in 228 African Americans from Connecticut (AACT) using Maximum Likelihood. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry.
In the two Latino groups EU ancestry was the highest, with substantial WAF and IA admixture. Both samples show wide variation in individual ancestry levels. Figure 2.3 shows distribution of genomic ancestry in the LCT sample. Average individual ancestry proportions in LCT is $0.19 \pm 0.16$ IA, $0.59 \pm 0.19$ EU and $0.22 \pm 0.18$ WAF. This result is similar to previously reported estimates of ancestry in other Latino populations (Bonilla et al. 2004a; Bonilla et al. 2004b). Only one individual shows 100% EU ancestry. 15.6% of individuals showed no IA ancestry and have only EU and WAF ancestry and 9.8% show no WAF ancestry and have EU and IA ancestry. Figure 2.4 shows distribution of proportional genomic ancestry in the LCA sample. In this sample proportional WAF ($0.11 \pm 0.14$) and IA ($0.36 \pm 0.18$) ancestry differ from that observed in LCT, with higher WAF ancestry in LCT and higher IA ancestry in LCA. Additional tests were performed to check for significant difference in individual ancestry levels between the two Latino populations (discussed later).
Figure 2.3: Distribution of Individual ancestry proportions in 102 Latinos from Connecticut (LCT) using Maximum Likelihood. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry.
Figure 2.4: Distribution of Individual ancestry proportions in 89 Latinos from California (LCA) using Maximum Likelihood. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry.
In addition to using MLE, a Bayesian MCMC method as implemented in the program STRUCTURE was also used to infer individual ancestry. The mean proportional individual ancestry estimates obtained with STRUCTURE are also shown in Table 2.2.

Triangle plots and associated histograms for EACT are shown in Figure 2.5, and those for AACT are shown in Figure 2.6.
Figure 2.5: Distribution of individual ancestry proportions in 652 European Americans from Connecticut (EACT) using STRUCTURE. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry.
Figure 2.6: Distribution of Individual ancestry proportions in 228 African Americans from Connecticut (AACT) using STRUCTURE. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry.
Plots and Histograms for the two Latino samples follow. The LCT sample is shown in Figure 2.7 and the LCA sample is shown in Figure 2.8
Figure 2.7: Distribution of Individual ancestry proportions in 102 Latinos from Connecticut (LCT) using STRUCTURE - IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry. Proportional EU
Figure 2.8: Distribution of Individual ancestry proportions in 89 Latinos from California (LCA) using STRUCTURE. IA: Indigenous American ancestry, WAF: West African ancestry and EU: European ancestry.
Significant correlation is observed in individual ancestry levels for estimates obtained with MLE and STRUCTURE as shown in Table 2.3, although mean individual ancestry estimates are different based on the method of estimation. Figure 2.9 shows the correlation in the estimates obtained using the two methods for the ancestral group contributing maximally to that population. It is notable that populations with known history of substantial admixture, such as the two Latino populations, show higher correlation between estimates obtained with the different methods. The histograms further show that ancestry estimates for the three admixed populations appear to be bound within certain limits of ancestry proportions, for example, using ML 35% of AACT show 0 IA ancestry, in contrast, using STRUCTURE, 35% AACT show 0.05 IA ancestry. It should be pointed out though, that not all individuals who have 0 IA ancestry with ML show 0.05 IA ancestry with STRUCTURE. A significant issue with the results obtained with STRUCTURE appears to be the tighter clustering of individuals from admixed populations and departure of estimates of AACT, LCT and LCA samples away from extreme edges of the plots, when compared to the estimates obtained with ML. While Bayesian methods such as the one implemented in STRUCTURE are modeled to account for uncertainty in the estimation process, ML assumes certain fixed parameters and makes inferences based only on the parameter estimates provided. Thus, using ML, one would expect to obtain extreme values of ancestry based on specific model assumptions while the Bayesian methods will try to adjust the data in a way that best fits the model. This could partly explain our observation. Further implications and significances for model choice are discussed later.
Table 2.3: Correlation between estimates obtained with ML and STRUCTURE

<table>
<thead>
<tr>
<th></th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>AACT</td>
<td>0.689</td>
<td>0.796</td>
<td>0.841</td>
</tr>
<tr>
<td>EACT</td>
<td>0.43</td>
<td>0.267</td>
<td>0.214</td>
</tr>
<tr>
<td>LCT</td>
<td>0.924</td>
<td>0.917</td>
<td>0.914</td>
</tr>
<tr>
<td>LCA</td>
<td>0.967</td>
<td>0.944</td>
<td>0.926</td>
</tr>
</tbody>
</table>

Shown are Spearman’s correlation coefficients. P values for all correlations were <0.0001

Figure 2.9: Correlation between estimates obtained with ML and STRUCTURE
**Effect of sample composition on estimates of ancestry**

The Bayesian MCMC method implemented in STRUCTURE simultaneously makes inferences for different parameters based on the total data available, which includes ancestral samples as well as the admixed population sample. Thus, estimates of ancestry obtained for any one individual depends on other individuals included in the analysis. To empirically evaluate the effect of sample composition, i.e. which precise individuals are included in the model, on ancestry estimates, 3 different random subsets of 100 EACT and 100 AACT individuals were chosen and analyzed independently using STRUCTURE. In each case the model was exactly the same as before, for the runs with all individuals, except that only 100 EACT or 100 AACT individuals were included in each run instead of the entire sample of 652 EACT or 225 AACT. Triangle plots of the individuals in each of the three runs, when included in the large sample or separately when included in the small sample are shown in Figure 2.10 for the EACT sample and in Figure 2.11 for the AACT sample. In both cases, all panels on the left show distribution of ancestry in subsets when using the entire set of individuals and all panels on the right show the subsets when using only the 100 individuals included in the analysis. Mean estimates of ancestry obtained when only 100 individuals are included in the run, compared to the estimates for the same individuals obtained previously when all individuals were included in the run are shown in Table 2.4 and in Table 2.5. As seen in these tables in the EACT, EU ancestry differed significantly in two out of three runs (Runs 1 and 2). In both cases, proportional EU ancestry decreases and proportional IA ancestry increases, when smaller numbers of individuals are included in the analysis. In Run 2, proportional WAF ancestry also decreased significantly (P<0.0002). Although these differences are significant, absolute values for the differences are small and the biological effect of these yet remains to be determined. In contrast, there were no significant differences in means for any of the ancestry axes in AACT when using a small sample vs. a large sample.
Table 2.4: Effect of sample size on Ancestry estimates in EACT using STRUCTURE

<table>
<thead>
<tr>
<th></th>
<th>Small Sets (N = 100)</th>
<th>Total Sample (N = 652)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WAF</td>
<td>EU</td>
</tr>
<tr>
<td>Run1</td>
<td>0.01 ± 0.01</td>
<td>0.96 ± 0.01</td>
</tr>
<tr>
<td>Run2</td>
<td>0.003 ± 0.001</td>
<td>0.96 ± 0.02</td>
</tr>
<tr>
<td>Run3</td>
<td>0.01 ± 0.06</td>
<td>0.98 ± 0.07</td>
</tr>
</tbody>
</table>

Shown are mean ± Standard Deviation for each ancestry axis. *Indicates significant difference between runs (P<0.0001 after Bonferroni correction for multiple testing) using all individuals to those using only subset of individuals.

Table 2.5: Effect of sample size on Ancestry estimates in AACT using STRUCTURE

<table>
<thead>
<tr>
<th></th>
<th>Small Sets (N = 100)</th>
<th>Total Sets (N = 228)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WAF</td>
<td>EU</td>
</tr>
<tr>
<td>Run1</td>
<td>0.79 ± 0.07</td>
<td>0.17 ± 0.07</td>
</tr>
<tr>
<td>Run2</td>
<td>0.81 ± 0.08</td>
<td>0.16 ± 0.07</td>
</tr>
<tr>
<td>Run3</td>
<td>0.79 ± 0.09</td>
<td>0.19 ± 0.08</td>
</tr>
</tbody>
</table>

Shown are mean ± Standard Deviation for each ancestry axis. 1 P = 0.004, 2: P<0.001: significant difference between runs using all individuals to those using only subset of individuals.
Figure 2.10: Comparing effect of large sample size (left panels) to small sample size (right panels) in EACT samples.
Figure 2.11: Comparing effect of large sample size (left panels) to small sample size (right panels) in AACT samples
Effect of model choice on ancestry estimates – Using European Americans as ancestors:

Ancestry estimates obtained with any method depend critically on the ancestral populations assumed to have contributed to the admixed sample. In our study we have used Spanish as a reference European ancestral population. While this choice is ideal for the two Latino samples, for the EACT and AACT samples Spanish is not the ideal European parental population. It is known that British, Irish, German and Italians contributed significantly to the European American gene pool. Therefore, we hypothesized that using the Spanish as an ancestral population may have some effect on the ancestry estimates. Although, this effect is likely to be of less consequence in comparison to using a single representative West African population, since variation within Africa is greater than that within Europe. We were unable to test the effect of using different African populations in this study. However, assuming that a random sample of European Americans are representative of different European populations, we tested the effects of model choice on ancestry estimates by substituting a panel of European Americans and estimating proportional ancestry of EACT (excluding those that are used as parentals) and AACT samples.

We selected a random panel of 59 EACT individuals and substituted these for the Spanish and then re-estimated ancestry for the AACT and EACT (except individuals used as ancestors who are referred to as EA henceforth) samples. Mean estimates for AACT using EA as ancestors is $0.83 \pm 0.08$ WAF, $0.14 \pm 0.06$ EA and $0.03 \pm 0.02$ IA with STRUCTURE and these were significantly different ($P<0.0001$ for IA, $P = 0.002$ for WAF and $P= 0.001$ for EA) from estimates obtained previously with EU ancestors (See Table 2.2). Mean estimates with ML are $0.69 \pm 0.25$ WAF, $0.23 \pm 0.24$ EA and $0.08 \pm 0.13$ IA. These estimates are significantly different for EA ($P=0.002$) and WAF (0.001) ancestry but not for IA ($P=0.108$) ancestry. Mean estimates for EACT (excluding those
that have been used as parentals) are $0.99 \pm 0.01$ EA, $0.004 \pm 0.01$ WAF and $0.01 \pm 0.004$ IA with STRUCTURE and $0.85 \pm 0.2$ EA, $0.07 \pm 0.14$ WAF, and $0.09 \pm 0.15$ IA with ML. The estimates with STRUCTURE were significantly different from those obtained previously ($P<0.0001$ in each case). The mean estimates with ML did not differ significantly from those obtained previously for the same set of individuals ($P=0.05$ for IA, $P=0.42$ for EA and $P=0.17$ for WAF). The choice of ancestral populations, especially with microsatellite markers has important consequences and is discussed later.

**Testing for difference between two Latino Populations:**

The two Latino groups collected from two different locations show different levels of genomic ancestry. As seen in Table 2.2, LCT has higher EU and WAF ancestry compared to LCA, while LCA has higher IA ancestry. We tested for significant difference in admixture levels using two different methods. We performed two-sample t-tests on admixture estimates obtained for each axis separately. This test was aimed at evaluating the significance of the difference in mean estimates obtained in one particular ancestry component. The results of the t-tests are shown in Table 2.6. Estimates of IA and WAF ancestry obtained with both ML and STRUCTURE shows significant difference between the two populations that persist after lowering the alpha level to 0.016 to account for multiple testing using Bonferroni correction. STRUCTURE shows significant difference between the two populations for all three ancestry axes.
Table 2.6: Comparing ancestry estimates in two Latino populations

<table>
<thead>
<tr>
<th>Ancestry</th>
<th>t-test</th>
<th>KS test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE</td>
<td>STR</td>
</tr>
<tr>
<td>EU</td>
<td>0.041</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>WAF</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>IA</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Table shows P values for pair wise comparisons. Significant level for each comparison is 0.01 after adjusting for multiple tests. Significant results in bold

We next compared the difference in individual ancestry estimates using the Kolmogorov-Smirnov test for two samples, which allows us to compare distributions of ancestry. We tested the hypothesis of no difference between the two data sets and rejected the null hypothesis for P values lower than 0.016 (after Bonferroni correction for multiple testing). Table 2.6 also shows the results of the K-S test. While significant differences were observed for proportional WAF (D = 0.36, P<0.0001) and IA (D = 0.55, P<0.0001), proportional EU did not show significant difference between the two groups (D = 0.15, P = 0.195), when using the ML. However, estimates obtained with STRUCTURE show significant difference for all ancestry components. Cumulative Fraction plots from K-S tests comparing the two Latino populations are shown in Figure 2.12. The X axis represents ancestral proportions and the Y axis represents cumulative ancestry. Solid lines represent LCT and broken lines represent LCA samples. Each step in the plot corresponds to a data point. The maximum vertical deviation is used as the test statistic for KS test. Although LCA has higher values for EU ancestry this is not significantly different from that observed in LCT. LCT has significantly higher proportional WAF ancestry compared to LCA. In contrast, LCA sample shows significantly higher IA ancestry than LCT sample. Taken together these tests indicate that there are differences in ancestry levels between the two groups. This was expected based upon differences in population history correlated with geographical sampling locations.
Figure 2.12: Cumulative Fraction plots from K-S tests comparing the two Latino populations.
**Tests for admixture stratification within samples:**

To test for presence of admixture stratification within the samples we used the individual ancestry correlation (IAC) test, where the original marker panel is randomly split into two separate panels of unlinked markers and individual ancestry is re-estimated using these marker sets. The procedure was repeated 20 times, to surmount effects of specific marker characteristics, (i.e. informativeness of marker panels in estimating individual ancestry). Based on previous simulation results, estimates obtained with the different marker panels are expected to be correlated in the presence of underlying stratification (Pfaff et al. 2001). Table 2.7 shows the results of the IAC test. While we detected evidence of population stratification in the two Latino populations and the AACT sample we failed to detect evidence of stratification in the EACT sample.

<table>
<thead>
<tr>
<th>Population sample</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EACT</td>
<td>0.041 (0.3)</td>
<td>0.067 (0.09)</td>
<td>-0.052 (0.187)</td>
</tr>
<tr>
<td>AACT</td>
<td>0.007 (0.917)</td>
<td>0.13 (0.05)</td>
<td>0.18 (0.006)</td>
</tr>
<tr>
<td>EC-CT</td>
<td>0.236 (0.017)</td>
<td>0.276 (0.005)</td>
<td>0.549 (&lt;0.0001)</td>
</tr>
<tr>
<td>WC-CT</td>
<td>0.289 (0.006)</td>
<td>0.404 (&lt;0.0001)</td>
<td>0.392 (&lt;0.0001)</td>
</tr>
</tbody>
</table>

Shown are Spearman’s correlation coefficient $\rho$ and associated P values. Significant P values in bold

**STRUCTURE** was set up to run each population separately for K=1, K=2 and K=3 separately. In each case STRUCTURE was used to test for the presence of subpopulations, as measured by deviations from HWE and linkage equilibrium in the sample, given the genotypes, by specifying models with predefined numbers of subpopulations (K=1, K=2 or K=3). In the AACT sample individuals were assigned to 2 subpopulations with greater probability than to 1 or 3 subpopulations. In all the other samples, EACT, LCA and LCT, K= 1 was the best option, indicating no evidence of
substructure within samples. Thus, using one test we found evidence for admixture stratification in all three populations with known history of admixture. Using the second test, AACT shows presence of stratification but neither of the Latino samples show substructure. Neither of the tests detected admixture stratification in the EACT sample.

**Phylogenetic analysis:**

To further investigate the relationship between the population groups a phylogenetic analysis of all populations was performed. Genetic distance between the populations was ascertained as pair wise $F_{ST}$ measures using Genepop software (Rousset and Raymond 1995). This distance matrix was subsequently used to construct a Neighbor-Joining tree of populations and is shown in Figure 2.13. As seen from this figure, the major divergence in observed between African and Non-African populations. EACT and EU samples cluster together as do the AACT and WAF samples. The two Latino populations clustered separately, the LCA clustering closer to the IA population, and LCT clustering closer to the WAF population, than the LCA sample. Branch lengths of the populations with recent history of admixture are shorter than those of the other non-admixed (in this case ancestral) populations. This further illustrates the difference between the two Latino samples and confirms the overall relationship between the different populations being studied.
Genetic distance between populations was measured as Fst, using the Genepop software. EACT: European Americans from Connecticut, EU: Spanish from Valencia, Spain, LCA: Latinos from California, LCT: Latinos from Connecticut, AACT: African Americans from Connecticut, WAF: Mende from Sierra Leone

Figure 2.13: Phylogenetic tree of all populations constructed using the Neighbor-Joining method implemented in MEGA 2

Simulation studies:

The panel of markers used for estimating individual ancestry includes the CODIS forensic markers, for which ancestry information content is typically low. However, with the additional high δc markers, total ancestry information content of the panel has been raised substantially (Yang et al. 2005). We performed additional simulation studies to evaluate the reliability of the estimates obtained with this panel. We simulated sets of individuals with 100% ancestry from one ancestral population using the EU, WAF and IA allele frequencies. Estimates of individual ancestry proportions were ascertained using ML method. These simulated individuals are expected to show 100% ancestry from the corresponding ancestral population. Table 2.8 shows the mean ancestry proportions of 1000 simulated individuals each from the EU, WAF and IA populations, estimated using
ML and STRUCTURE respectively. Theoretically, a homogeneous unadmixed population would be expected to cluster tightly at one vertex, with minimal dispersion of individual ancestry proportions along any of the ancestral axes of the triangle plot. In the simulated sample, the proportion of individuals with 100% ancestry from the respective ancestral population was found to be 63.6% for IA, 57% for EU and 57.7% for WAF. Based on these results there appears to be a bias (as measured by proportional ancestry from populations that did not contribute to the simulated samples) of 3% for IA, 4% for WAF and 4% for EU in admixture estimates obtained using this marker panel. Bias in ancestry estimates for a population - “population bias”, is calculated as the total ancestry from the non-contributing populations, for example, total non-EU ancestry measured in the simulated EU samples. This estimate does not reflect whether such bias can be attributed to one specific non-contributing ancestral population (where more than two ancestral populations are included in the model). In the example above population bias represents total non-EU ancestry irrespective of whether it is WAF or IA ancestry measured in the simulated EU samples. This measure can be used to evaluate the reliability of observed estimates in the study samples. A second measure of bias is the “ancestry bias” which is complimentary to population bias in that this is a measure of the total contribution from one non-contributing population to other populations in the analysis. For example the total IA ancestry in simulated 100% EU individuals or the total WAF ancestry in simulated 100% IA individuals. The population bias for all ancestral groups was less than 4% in this sample, when using ML, indicating that there is up to 4% chance of individuals showing ancestry from a population from which they have no admixture, just by chance, based upon the estimates obtained with the current marker panel. Ancestry bias for both IA and EU are up to 4% on average. Only 1% ancestry bias was observed on average for WAF, indicating that using this marker panel there is less chance of detecting some African ancestry where none is present. These estimates further demonstrate that this marker set is less able to distinguish between IA and EU, compared to WAF and non-WAF distinctions, a finding that is consistent with the expectations based on the procedure used to select the markers (Yang et al. 2005). With STRUCTURE the bias estimates are less than 1% in each case, indicating that greater efficiency is
achieved in estimating ancestry proportion for homogeneously non-admixed samples using the present marker panel and the STRUCTURE algorithm, when a sufficiently large data set is used.

Table 2.8: Estimated ancestry in simulated sample of unadmixed individuals

<table>
<thead>
<tr>
<th>Population</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
<th>Population Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IA</td>
<td>0.98 ± 0.05</td>
<td>0.02 ± 0.05</td>
<td>0.004 ± 0.02</td>
<td>0.03 ± 0.07</td>
</tr>
<tr>
<td>AF</td>
<td>0.02 ± 0.03</td>
<td>0.02 ± 0.04</td>
<td>0.97 ± 0.05</td>
<td>0.04 ± 0.07</td>
</tr>
<tr>
<td>EU</td>
<td>0.03 ± 0.05</td>
<td>0.97 ± 0.06</td>
<td>0.01 ± 0.03</td>
<td>0.04 ± 0.08</td>
</tr>
<tr>
<td>Ancestry Bias</td>
<td>0.04 ± 0.08</td>
<td>0.04 ± 0.08</td>
<td>0.01 ± 0.04</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IA</td>
<td>0.99 ± 0.004</td>
<td>0.002 ± 0.002</td>
<td>0.003 ± 0.003</td>
<td>0.05 ± 0.05</td>
</tr>
<tr>
<td>AF</td>
<td>0.003 ± 0.003</td>
<td>0.003 ± 0.003</td>
<td>0.99 ± 0.01</td>
<td>0.05 ± 0.05</td>
</tr>
<tr>
<td>EU</td>
<td>0.002 ± 0.001</td>
<td>0.99 ± 0.001</td>
<td>0.001 ± 0.001</td>
<td>0.003 ± 0.002</td>
</tr>
<tr>
<td>Ancestry Bias</td>
<td>0.005 ± 0.004</td>
<td>0.005 ± 0.005</td>
<td>0.004 ± 0.004</td>
<td></td>
</tr>
</tbody>
</table>

Although ancestry estimates obtained with ML and STRUCTURE show high correlation, there appear some discrepancies between these estimates. To explore these differences further, we simulated a sample of individuals based on the proportional ancestry of the 228 African Americans in our sample using the current marker panel. Ancestry proportions in the simulated sample were made to vary corresponding to the MLE ancestry proportions observed in the AACT sample. We simulated 14 sets of individuals with mean ancestry proportions as shown in Table 2.9. The individuals were simulated for varying numbers of generations and different admixture proportions. For instance, to simulate 6 individuals with 0.96 WAF and 0.04 IA, we started with 61 WAF ancestors and 4 IA ancestors 6 generations in the past. The resulting sets of simulated
individuals were analyzed using both ML and STRUCTURE and the results are shown in Table 2.11. Pair wise t-tests comparing ML and STRUCTURE estimates to expected ancestry show non-significant differences for MLE, while significant differences are seen between STRUCTURE and expected estimates for WAF and IA ancestry, but not for EU ancestry. Spearman’s correlations between expected and observed estimates are shown in Table 2.10. Both methods show nearly equivalent correlation separately with the expected estimates. Cumulative fraction plots of expected WAF vs. estimated WAF ancestry using the two methods are shown in Figure 2.14. It is interesting that with ML extreme values of ancestry are obtained at the lower end of the distribution where none is expected, while with STRUCTURE the values obtained at the lower end of the distribution are higher than expected.
X axis shows proportional WAF ancestry in the sample. Y axis shows cumulative fraction. Solid line indicates the expected ancestry estimates in simulated sample. Broken lines indicate genomic ancestry measure with the two different methods.

Figure 2.14: Cumulative WAF ancestry distribution in simulated samples using ML (shown in top panel) and using STRUCTURE (shown in bottom panel)
### Table 2.9:  Ancestry composition in simulated samples

<table>
<thead>
<tr>
<th>No. of individuals</th>
<th>AF</th>
<th>EU</th>
<th>IA</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.96</td>
<td>0</td>
<td>0.04</td>
</tr>
<tr>
<td>27</td>
<td>0.88</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>12</td>
<td>0.85</td>
<td>0</td>
<td>0.15</td>
</tr>
<tr>
<td>35</td>
<td>0.83</td>
<td>0.02</td>
<td>0.15</td>
</tr>
<tr>
<td>25</td>
<td>0.83</td>
<td>0.04</td>
<td>0.13</td>
</tr>
<tr>
<td>15</td>
<td>0.8</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.8</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>25</td>
<td>0.73</td>
<td>0.1</td>
<td>0.17</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>40</td>
<td>0.68</td>
<td>0.08</td>
<td>0.24</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.4</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.35</td>
<td>0.45</td>
<td>0</td>
</tr>
</tbody>
</table>

First column indicates number of samples in each group. Groups ordered from highest to lowest proportional WAF ancestry.

### Table 2.10:  Mean estimates of ancestry obtained with MLE and STRUCTURE in simulated sample

<table>
<thead>
<tr>
<th></th>
<th>MLE-IA</th>
<th>MLE-EU</th>
<th>MLE-AF</th>
<th>ST-IA</th>
<th>ST-EU</th>
<th>ST-AF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.15 ± 0.07</td>
<td>0.14 ± 0.15</td>
<td>0.8 ± 0.15</td>
<td>0.03 ± 0.02</td>
<td>0.16 ± 0.09</td>
<td>0.82 ± 0.09</td>
</tr>
<tr>
<td>P value</td>
<td>(0.716)</td>
<td>(0.334)</td>
<td>(0.419)</td>
<td>(&gt;0.0005)</td>
<td>(0.719)</td>
<td>(0.012)</td>
</tr>
</tbody>
</table>

P values of pair wise t-tests comparing expected and observed estimates shown in parenthesis. Significant differences are in bold.
Table 2.11: Correlation in estimates of simulated individuals for ML and STRUCTURE

<table>
<thead>
<tr>
<th></th>
<th>IA</th>
<th>WAF</th>
<th>EU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML</td>
<td>ST</td>
<td>ML</td>
</tr>
<tr>
<td>Expected</td>
<td>0.49</td>
<td>0.47</td>
<td>0.753</td>
</tr>
<tr>
<td>ML</td>
<td>0.63</td>
<td>0.83</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Shown are Spearman’s ρ. P Values <0.0001 in each case

Discussion:

Using a panel of 37 STR and SNP markers, we have estimated individual ancestry in different populations and observed variation in ancestry estimates in all populations. Variations in individual ancestry estimates were observed as expected in the Latino and AACT populations. We also observed some variation in individual admixture in the EACT population samples.

Microsatellite panel for ascertaining individual ancestry

The marker panel used in the present study is comparable to AIM panels used previously for individual ancestry estimation in terms of information content as measured by total $\delta_c$ (Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004). For instance, Bonilla et al., 2004a used a panel of 36 AIMs where total $\delta$ was 11.72 for EU-IA comparison, 14.28 for EU-WAF comparison and 16.34 between WAF-IA (Bonilla et al., 2004a). In comparison, total $\delta_c$ in the present panel is 12.6 for EU-IA comparison, 14.35 for EU-WAF comparison, and 17.3 for WAF-IA comparison. Total information content of the present marker panel consisting mostly of STRs, is comparable to that of SNP...
panels used in the previous studies (Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004; Reiner et al. 2005) and thus provides an alternative panel for individual ancestry estimation of comparable power. Individually, more than 50% of the marker panel is informative for any pair-wise comparison between populations as indicated by the individual marker $\delta_c$ values ($\delta_c > 0.3$). In addition, we report allele frequencies for all markers in a South East Asian (Thai) population. The choice of markers has been optimized for easy amplification and genotyping (as discussed in Yang et al., 2005) and promises to be a good resource for ancestry estimation.

Previous studies have indicated that markers that distinguish between AS and EU populations are likely to be as informative for EU-IA comparisons (Collins-Schramm et al. 2004). In the marker panel used for the present study we observed small but significant correlations between the EU-IA and EU-AS $\delta_c$ values, as well as WAF-IA and WAF-AS $\delta_c$ values. While this can be interpreted as evidence for using markers with high AS-EU $\delta_c$ values to screen for markers that distinguish between IA and EU populations, it is also possible that these markers actually distinguish between EU and other non-EU populations generally. In cases where more than one non-WAF, non-EU population has contributed to the admixed group, such markers may be less useful, but we were not able to evaluate this hypothesis empirically. The total $\delta_c$ value for the present panel of markers between AS-IA populations is intermediate between that in EU-AS and EU-IA populations. Taken together this probably reflects the shared recent common ancestor between the three populations. In relation to other pair wise comparisons there appears to be some information in this panel of markers to distinguish among different non-WAF populations, although to efficiently estimate individual ancestry with a standard error of no more than 0.1 we would require at least 40 unlinked AIMs with average ancestry information content “$\delta$” of 0.67 (McKeigue et al. 2000). As the analysis of the EACT sample indicates, in this marker panel lower discrimination power between non-European samples may be problematic. Further efforts should be made to identify a more balanced panel of markers that distinguish equally between all population samples included in the model.
Estimates of individual ancestry in different populations

We have estimated individual ancestry proportions in major resident US populations using two different methods. Correlation between STRUCTURE and ML estimates were highest in the Latino populations which are also more admixed. Spearman’s correlation coefficient > 0.9 for all ancestry axes with associated P values < 0.0001. We observed the least correlation in EACT population. STRUCTURE estimates showed tighter clustering of EACT individuals compared to the ML estimates (Figures 2.1 and 2.5). Although there was less correlation in the EACT estimates between estimates obtained with the two methods (0.43 for IA, 0.27 for EU and 0.21 for WAF, P<0.0001 in all cases) there is observable variation within the sample, with some individuals showing higher IA admixture while some with higher WAF admixture as shown by the triangle plots. However, it should be noted that in the triangle plots individuals with the same ancestry proportion all plot together and one point on the plot may represent one or even one hundred individuals. While the triangle plots provide a means of visual representation of variation in the sample, they have a negative aspect in that they do not indicate what proportion of the samples each point on the plot represents. So, the plots for EACT need to be evaluated cautiously, especially when being compared against the plots of AACT, LCT, and LCA. In addition, the sample size for EACT is by far the largest in this study (652) compared to that of the other populations. So, informally speaking, while the variation in the EACT sample using ML appears to be quite high, the proportion of individuals who are responsible for this variation are few. This is more clearly indicated in the histograms showing estimates obtained from ML and STRUCTURE. A related issue is the difference in estimates obtained with ML and STRUCTURE, which is more pronounced in the EACT sample than in other admixed samples. Some of this discrepancy is explained by the difference in the methods by which ML and STRUCTURE make inferences about ancestry proportions and are discussed in the next section. Nevertheless, results from both methods indicate that there is some variation among the EACT that may be important to consider when studying this population. Further tests and simulation studies with more informative marker panels are
required to better understand the variation within EACT populations. The tests of stratification did not detect any genetic substructure in the EACT sample. It is possible that larger panel of ancestry informative markers is required to detect admixture stratification within EACT, as has been recently shown using a panel of 11,555 SNP markers (Shriver et al. 2005). In addition, it should be remembered that the markers were chosen do distinguish between European and African populations. Therefore, it is likely that these markers are suppressing variation within European populations and to understand stratification within European American samples we will require different sets of markers that specifically distinguish between European populations. The lower discrimination power between EU-IA, compared to EU-WAF and IA-WAF distinction, by the present marker panel may be responsible to some extent for the elevated proportional IA ancestry in this EACT sample. Using a marker panel that provides equal discrimination power between all populations is likely to give more information regarding genetic stratification in European Americans.

The estimates of individual ancestry in the AACT and Latino samples are consistent with estimates reported previously (Parra et al. 1998; Bonilla et al. 2004). Since African American and Latino populations vary depending on where they are sampled (reflecting differences in population history and social norms), we did not expect to obtain exactly the same estimates in these populations as those reported previously. We have analyzed the AACT samples under a three-way admixture model, since it is possible that some individuals in this population have IA admixture. In the present marker panel African-non African $\delta_c$ is the highest and therefore it is easier to distinguish someone who has high non-African ancestry. In addition, the IAC test showed significant correlation between the estimates of ancestry obtained with different marker panels (See table 2.7). Tests with STRUCTURE also detected presence of admixture stratification in this sample.

We observed significant differences between the two Latino populations with the Latinos from Connecticut showing higher WAF ancestry and those from California showing higher IA ancestry as shown in Table 2.2 and Figure 2.12. The results for variation in IA ancestry are similar to those of a previous report (Bertoni et al. 2003).
However, in contrast to their observation of similar WAF ancestry across Latino groups from different locations in the US, we observed significantly higher WAF ancestry in the population from Connecticut. This possibly reflects the demographic history of these populations and it is likely that the sample from California is predominantly Mexican American (US Census, 2000), with higher IA ancestry, while the sample from Connecticut is probably Cuban or Puerto Rican (US Census, 2000) and has higher WAF admixture.

The Phylogenetic analysis further reflects the relationships between the populations. As seen in Figure 2.13, populations of primarily African ancestry clearly distinguish from those of non-African ancestry, with the two samples of European origin, one Spanish (EU) and the other European Americans from Connecticut (EACT) clustering very close to each other with small branch lengths. The two populations of African origin, African Americans (AA) and the Mende (WAF) cluster together. The African Americans have a short branch length, which shows that though they are closer to the Mende, they have some divergence and are admixed. The Indigenous Americans (IA) is more closely related to the European samples but show longer divergence from both African and European populations. The two Latino samples clearly cluster separately; with the LCT clustering more closely to the African branch and the LCA clustering with the Indigenous American branch. This further illustrates the difference between the Latino samples.

In both Latino samples, evidence of admixture stratification was detected only with IAC test and not with STRUCTURE. Given the known history of admixture in these populations it is likely that a continuum in ancestry proportions exists, as opposed to distinct subgroups. Indeed, estimates of individual ancestry show this, using both ML and STRUCTURE. Variation in individual ancestry levels is an indicator of possible confounding in case-control associations if not taken into consideration, since disease risk may vary with proportional ancestry. The two methods used in this analysis have some differences that should be clarified here. STRUCTURE attempts to identify subgroups based on deviations from HWE and linkage equilibrium in the model. It is possible that random mating within the population has resulted in a situation where neither of the
assumptions of the STRUCTURE algorithm is violated and hence, the program detects no evidence for subpopulations. The assumptions in IAC test are different. Here we test for relative differences between individuals in the population. If ancestry proportions vary substantially within the population, then on average all regions of the genomes in each individual will reflect this variation. Thus, we can look at any part of the genome and get an estimate of ancestry for an individual. Across individuals, the relative differences will be retained, no matter which part of the genome is used to infer ancestry proportions for each individual. Simulation studies have shown that in the absence of admixture stratification no association is observed between ancestry proportions measured with different sets of markers (Pfaff et al. 2001).

While correlation in ancestry estimates obtained with different methods may be used as an indicator of reliability of these estimates (Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004; Reiner et al. 2005; Shriver et al. 2005), it is interesting to compare situations where the estimates obtained with the two methods differ. Even though correlations are strong in the AACT, LCT and LCA samples, the correlation plots in Figure 2.9 indicate some discrepancies in estimates obtained with the two methods. This is especially evident in the EACT sample. Ideally, the two estimates would closely approximate each other with the fitted line passing through zero and having a slope of 1. This issue was explored further through simulations and is discussed below.

Effects of models, methods and markers on ancestry estimates:

The main focus of the present study was to ascertain ancestry proportions in different US populations using a marker panel, which has sufficient information. This was done to obtain a description of underlying variability within populations. Previous studies on admixture have typically focused on a single population (e.g., African Americans from different US locations by Parra et al., 1999, Puerto Ricans by Bonilla et al., 2004a, Hispanics from Colorado by Bonilla et al., 2004b, African Americans from four US cities by Reiner et al., 2005). Analyzing different populations using the same
marker panel and a using a reasonable admixture model based on known population history of the samples, allowed us an opportunity to identify certain issues that should be considered for ancestry estimation. The estimates of ancestry obtained for AACT and Latino samples are similar to those reported in literature. However, estimates of ancestry in the EACT sample were interesting. While it may seem inappropriate at a first glance, to analyze individual ancestry in European Americans using a three way admixture model as we have done here, given the knowledge of peopling of the Americas it is highly likely that some very low levels of non-European ancestry is likely to exist in the larger European American population. Given the close association between the three ancestral populations in the US for the last five hundred years, this is inevitable. Further, the subjects sampled were self identified as European Americans and verified by interviewer’s observation. Self-identification is often a cultural process with very little consideration of underlying genetic background. While in African American and Latino populations, the large-scale history of migration and admixture is a well documented and a continuing process, it is largely assumed the European Americans are more homogeneous group. However, we have used two different methods and both show that there is evidence of low levels of non-European admixture in this sample. The biological significance of this variation or the possible effects that this may have on association studies warrants further investigation.

In our model of admixture we have used Spanish as a representative European ancestral population. While this sample is well suited for analyzing the two Latino samples, in case of AACT and EACT, there are likely to be some consequences for using this as a parental population. It is known that the current European American population has ancestry from several European ancestral populations including British, Irish, Germans and Italians. A more appropriate ancestral population would be one where we could include several extant European population samples. Variation within the African continent is perhaps more important to consider than variation within Europe which is why we have focused on a West African ancestral population. Considerations for model choice are illustrated by our analysis of EACT and AACT samples using a subset of EACT as ancestral as an alternative of the Spanish sample. In case of the AACT samples,
both methods showed significantly different estimates compared to those obtained previously with Spanish as an ancestral population, however, whether these differences have any biological significance is yet to be determined. In the EACT sample, ML estimates did not differ significantly from those obtained previously, but estimates with STRUCTURE did, even though the absolute value of differences is small.

Part of the differences can be explained by the differences in allele frequencies in the Spanish and EACT samples. Total $\delta_c$ between the Spanish and EACT is 3.1, and 50% of the markers show greater than 10% allele frequency difference between the two populations. This results in changes observed in ancestry estimates with the different samples. Using the subset of EACT resulted in higher EU ancestry in the EACT sample for both ML (Average EU with Spanish ancestors is 0.84 ± 0.12 and with EACT ancestors is 0.85 ± 0.2) and STRUCTURE (Average EU with Spanish ancestors is 0.98 ± 0.02 and with EACT ancestors is 0.99 ± 0.01), but for the AACT sample EU ancestry was estimated to be higher with ML (Average EU with Spanish ancestors is 0.17 ± 0.14 and with EACT ancestors is 0.23 ± 0.24) but not with STRUCTURE (Average EU with Spanish ancestors is 0.16 ± 0.07 and with EACT ancestors is 0.14 ± 0.06). It was not possible to test which ancestral population provides more reliable estimates following the methods suggested by (Long 1991), and subsequently used by Bonilla et al., (2004b) where frequency difference between admixed population and ancestral populations are tested. Chiefly, because we have used microsatellites and in many cases alleles present in one population are not always present in another population.

It is important to briefly discuss the difference between the two approaches that have been used in this study. Both these methods have certain assumptions inherent in them and all results should be interpreted after considering these assumptions. ML relies on the exact ancestral population allele frequencies. More precise estimates of ancestry are expected if 1) appropriate ancestral populations are specified, 2) large sample sizes are used for estimating ancestral allele frequencies, 3) a complete data set is used which has few missing genotypes, 4) large panel of highly ancestry informative markers are used. In this method, estimates for each individual depend on the exact markers genotyped in that individual. In addition, since each locus is treated as an independent
observation, the information inherent in linked loci cannot be used explicitly, unless haplotypes are modeled and haplotype frequencies are used instead of allele frequencies for those loci that are tightly linked. The sample size of the admixed population under investigation has no effect on the estimate of any one individual. Thus, estimates of ancestry obtained with ML are reliable considering that all model assumptions have been met adequately.

In contrast, the Bayesian MCMC methods have been proposed to take into account the inherent uncertainty in 1) choice of ancestral populations i.e. the admixture model assumed 2) allele frequency estimates in ancestral populations and 3) ambiguities arising due to missing data. The STRUCTURE algorithm uses admixed and unadmixed individuals to make inferences for any one individual. This is clearly seen in the studies of sample sizes where subsets of EACT and AACT individuals have been used. The EACT sample is by far the largest in this dataset and is in far excess of the sample sizes of the three ancestral populations. The algorithm in STRUCTURE simultaneously infers allele frequencies in all populations and ancestry proportions in all individuals in the sample. We have included ancestral populations and used the “Usepopinfo” for augmenting the inference procedure by designating individuals that can be used for training the data for better updating the program. Thus, in the Bayesian framework the large EACT sample provides additional information that is used for inferring ancestry of each individual. It should be noted that in our study of subsets of 100 EACT and 100 AACT, mean estimates for the two groups did not change. However, as the triangle plots in Figure 4a show that for some EACT individuals, using the smaller sample gives very different estimates of ancestry than when the large sample is used. The variation seen within EACT is typically lower than that seen in AACT or Latino samples and could be due to few individuals included in the EACT sample who are genetically more diverse than the other individuals in this sample. This is also evident in our simulation of unadmixed samples. We simulated 1000 unadmixed individuals and tested against our model. Typically, the bias with STRUCTURE was lower than that with ML. It is speculated that when true variation among individuals in the sample being studied is low, the effect of a few individuals who are genetically distant from the majority of
individuals is minimized in a larger sample compared to a smaller sample. When the sample size is small, including a few individuals who are very different from the others will have a more appreciable effect. Further studies are required to confirm this assumption. However, as seen in the AACT sample sets, when there is greater variation in the sample, in this case due to substantially historically documented admixture in the sample, the sample size of the admixed population is less of an issue.

In the three admixed samples, STRUCTURE appears to cluster individuals more tightly, than does ML. Despite lower variances in ancestry estimates in admixed samples, correlation between the estimates is high. This could indeed reflect true ancestry proportions in the sample, or as in case of MLE, reflect an error in the underlying assumption. It is known that with large sample sizes and large panel of AIMs, both Bayesian and ML estimates will be asymptotically equivalent (McKeigue et al. 2000). Although we have relatively large numbers of individuals, there are only a few markers being used in this study. Since we have seen evidence of convergence, what we observe are the best estimates possible, given the data. Besides, the STRUCTURE algorithm assumes a Dirichlet prior distribution, which is a unimodal distribution, for allele frequencies and ancestry estimates. If the true distribution of these parameters in the sample were different, then the fit of the model would indeed be poor. This could be the case with both real and simulated samples.

The simulation study helps to clarify some basic discrepancies between the estimates obtained with the two methods. There is rarely a “Gold Standard” for comparing estimates of ancestry obtained with different methods. We have used expected ancestry (genealogical ancestry) based on the number of ancestral individuals from different ancestral populations, against which to compare estimates obtained with different methods.

The simulation of unadmixed individuals indicates that the present marker panel has sufficient power to distinguish between the three ancestral populations, using either method. In the simulated samples variation in individual admixture in the West African ancestral sample was of the same magnitude along both of the non-African axes, whereas in the European and Indigenous American samples variation was greater along the EU-IA
axis, possibly reflecting the more recent shared common ancestry between these populations and lower total $\delta_c$. While the large sample size may have affected the STRUCTURE estimates, which shows >99% ancestry for each ancestral population, ML which is unaffected by sample size shows that there is <5% chance of ancestry bias or population bias using this marker panel. These studies also provide an estimate of the level of bias expected using this marker panel for the two methods used in the study. Theoretically, for samples of individuals simulated to have no admixture, all individuals are expected to show 100% ancestry from one population. Variations observed depend on the ancestry information content of the marker panel and the method used for making inferences. For a set of 90 SNP markers in which different alternate alleles are fixed in different populations, ancestry estimates are observed as expected (IH unpublished work), when all individuals who are being studied are from one population and have no history of admixture. However, the current panel of STR is not one with such definitive ancestry information content. In our simulation study the estimates obtained with ML were not significantly different from expected estimates but estimates with STRUCTURE were significantly different for at least one major axis of ancestry. Despite difference, correlations were high between estimates obtained with the two methods. In addition, simulation of the admixed samples indicates that there is greater concordance between estimates obtained with different methods than that with the expected estimate. This finding is consistent with the random variation in the segregation of alleles at each generation, which would lead to variation in ancestry among individuals each having the same expected ancestry. Since our expectation was based on genealogical ancestry of individuals, there is likely to be variation in realized or genomic ancestry proportions and that is what each of the methods is measuring.

The choice of markers also has important effects on ancestry estimates. We have used a panel of microsatellites for estimating ancestry. Based on computed $\delta_c$ levels, this panel compares to panel of SNPs reported previously. However, it should be pointed that this measure consolidates information across multiple loci that can have important consequences. When di-allelic SNPs are used, assigning ancestry to an allele is easier, especially for AIMs since frequencies of only two possible alleles have to be considered.
In contrast, for multiallelic loci there is greater chance that some alleles will be missing in ancestral populations that exist in moderate frequency in the admixed sample, due to stochastic variation. MLE will treat such alleles as missing. In such cases it may be more efficient to use linked markers in Bayesian MCMC methods (like STRUCTURE or methods proposed by McKeigue et al. 2000) which can incorporate information in linked markers for making inferences of individual ancestry. Thus, choice of models and markers are both important when analyzing ancestry proportions. In the event that information is not available regarding admixture history, some reasonable assumptions will have to be made.

Obtaining reliable estimates of ancestry is always challenging. In most instances where estimates of ancestry have been obtained and used for further analysis it has been implicitly assumed that these estimates of ancestry are indeed reliable and invariant. However, it should be pointed that estimates of ancestry depend critically on model assumptions, markers used and as this study further demonstrates, methods used for this purpose, as we have discussed above. Taking into account considerations for marker choice, model assumption and methods used, empirical data and the simulations indicate that estimates of individual ancestry obtained with this panel provide a measure of underlying genetic variation in the samples.

In conclusion, this study provides a description of distribution of individual genomic ancestry in different populations. It shows that individual genomic ancestry varies in all populations, more in ones with known recent history of admixture. Such variation may not lead to the creation of distinct subgroups, but rather a continuum of individual admixture proportions that is likely to affect case-control associations unless accounted for. Comparisons of the two Latino groups further suggest that there is geographic heterogeneity among samples that reflects the demographic histories of the populations. In addition, using real and simulated data we have shown that certain considerations should be made regarding sample size, model choice and method of estimation when performing such analyses.
References:


Chapter 3

Admixture structure, Ancestry-Phenotype associations and Admixture Mapping of Hypertension and Obesity related phenotypes in two samples of African Americans from Philadelphia and Bogalusa

Introduction

Epidemiological studies indicate that several common complex diseases show differences in prevalence among ethnically classified groups (see table 1.1, Chapter 1). Some of these diseases are leading causes of morbidity and mortality and include type 2 diabetes, hypertension, obesity, coronary disease, and prostate cancer. Studies have shown that hypertension and prostate cancer are more common in African Americans (Gaines and Burke 1995; Douglas et al. 1996; Thompson et al. 2001; Hoffman 2002; Kramer et al. 2004), obesity (Hodge and Zimmet 1994) and non-insulin dependent diabetes mellitus is more prevalent in Hispanics and Indigenous Americans (Chakraborty and Weiss 1986; Knowler et al. 1988), in comparison to Europeans and European-American populations. Type 2 diabetes shows extensive ethnic variation in disease risk, for instance, compared to low-risk Europeans, high-risk groups include Indigenous Americans (Knowler et al. 1978), Pacific islanders (Zimmet et al. 1977), Indigenous Australians (Wise et al. 1976), South Asians (McKeigue et al. 1991) and Peninsular Arabs (al-Mahroos and McKeigue 1998). Family based linkage techniques are unlikely to be informative in such cases due to age dependent onset of the disease and various gene-environment interactions that also influence the trait and the problems of obtaining genetic data from multigenerational pedigrees. Association mapping studies that are designed to detect non-random occurrences of a marker allele with respect to a phenotype (e.g. affected disease status) have little power when disease susceptibility is
influenced by many rare alleles or allelic heterogeneity is high. Consequently, both
traditional family based designs and association approaches for mapping genes has had
limited success in identifying genes that cause or influence these phenotypes.

Association studies have received a lot of attention as means for mapping genes
that influence complex diseases (Risch and Merikangas 1996; Jorde 2000; Nordborg and
Tavare 2002; Schulze and McMahon 2002). These studies primarily look for non-random
associations of marker alleles with respect to a phenotype which result from either direct
biological effects of the marker alleles or, more often, from linkage disequilibrium (LD)
between the marker allele and a nearby predisposing or causal locus. The extent of LD is
a complex function of a number of genetic and evolutionary factors such as mutation,
recombination and gene conversion rates, demographic and selective events, and the age
of the mutation itself. Some of these factors affect the whole genome while others only
affect particular genome regions. Additionally, variation in mutation, recombination and
gene conversion rates throughout the genome are expected to create LD differences
between genomic regions (for example see (Taillon-Miller et al. 2000). Population choice
for LD mapping has prompted much discussion and debate (Wright et al. 1999; Eaves et
al. 2000; Kaessmann et al. 2002; Nordborg and Tavare 2002). Small, isolated and inbred
populations have been proposed by some authors, to be better suited for gene mapping,
than other populations, due to the lower heterogeneity and the larger extent of LD
(Wright et al. 1999; Kaessmann et al. 2002; Nordborg and Tavare 2002).

Admixture Mapping (AM) is a method that utilizes the LD created when two
ethnically distinct populations admix (as in US populations of mixed ancestry like
African Americans or Mexican Americans) and is well suited for complex disease gene
mapping. When previously separated populations come together creating a new admixed
population, non-random associations are generated at all loci, linked and unlinked where
allele frequencies are different among ancestral populations. The magnitude of admixture
linkage disequilibrium (ALD) is proportional to the allele frequency differential (δ)
between the parental populations for each locus under consideration. Admixture mapping
is appropriate for mapping complex disease genes in which the two parental populations
exhibit different levels of disease prevalence (Chakraborty and Weiss 1988; McKeigue
1997; McKeigue 1998) since the surrounding causal alleles are most likely in LD with the disease gene. The exponential relationship between decrease in LD and genetic distance facilitates the discrimination between LD in an admixed population, which remains high as markers are close together and genetically linked, and the background LD at unlinked loci. Critical parameters that influence admixture mapping are the frequency differential ($\delta$) of markers between the ancestral populations, the amount of admixture ($m$) and the number of generations since the admixture event. The statistical basis of this approach, originally investigated by Chakraborty and Weiss (1988) and subsequently by (Briscoe et al. 1994; Stephens et al. 1994) was designated as MALD (Mapping by admixture linkage disequilibrium). While ALD is extremely useful in that non-random associations are generated over long regions, admixture also causes stratification in a population through inter-individual variation in admixture in a population. The amount of structure introduced by admixture depends on several factors, including the level of admixture, dynamics of the admixture process, and social structure in the population that is related to ancestry (Pfaff et al. 2001). It is important to control for structure in the population in order to avoid false positive and false negative results (Parra et al. 1998; Lautenberger et al. 2000; Pfaff et al. 2001; Kittles et al. 2002; Nordborg and Tavare 2002; Hoggart et al. 2003). Several other methods have been proposed to control for population structure (Devlin and Roeder 1999; Pritchard et al. 2000; Hoggart et al. 2003). One way of controlling for spurious associations is to use estimates of proportional ancestry as a conditioning variable in association studies. Additionally, estimates of BioGeographical ancestry (BGA), the proportional genomic ancestry levels of an individual, can be used in conjunction with measured environmental effects for investigating the roles of environmental and inherited risks underlying complex traits (Fernandez et al. 2003; Gower et al. 2003; Molokhia et al. 2003). It is important to recognize that associations between individual admixture and disease risk might reflect correlations between BGA and socio-cultural variables and exposures. For example, hypothetically, if BGA and years of education were to be correlated, hypertension might be correlated with BGA, even though the causal risk factor were years of education or vice versa.
An alternative approach for exploiting admixture has been developed that is different from classical linkage disequilibrium mapping, and is more similar to linkage analysis of an experimental cross (McKeigue et al. 2000; McKeigue 2005) and hence is referred to as admixture mapping. In this method, the underlying variation in ancestry on chromosomes of mixed descent is modeled, to extract all the information about linkage that is generated by admixture. Despite the requirements of advanced statistical methods for application of this approach in practice, the underlying principle on which it relies to detect linkage is straightforward. Assume that a locus accounts for some of the variation in a phenotype between two ancestral populations (for example Europeans and West Africans). If individuals of mixed descent (i.e. those that have ancestry from both the parental populations) can be classified into whether they have 0, 1 or 2 alleles of West African ancestry, then it will be possible to test for the variation in the phenotype in these three groups, holding all other factors constant. It is possible to control for parental admixture (genealogical ancestry) in this analysis to eliminate association of the trait with unlinked loci. Ancestry of an allele at a locus is deduced by using the conditional probability of each allelic state given the ancestry of the allele (ancestry specific allele frequencies). Using a complex hierarchical model with many nuisance parameters it is possible to model the distribution of admixture in the population, the admixture of each individual’s parents, and the stochastic variation of ancestry on chromosomes inherited from these parents. Fitting such models requires a Bayesian approach with Markov chain simulation as previously described (McKeigue 1997; McKeigue 1998; McKeigue et al. 2000; Hoggart et al. 2003; Hoggart et al. 2004; McKeigue 2005). This method is implemented in the program ADMIXMAP (for Windows and Linux platforms) and is freely available at www.ucd.ie/genepi/softwatre/html.

Growing evidence suggests admixture mapping will be an effective means of gene identification (Shriver et al. 2003; Bonilla et al. 2004a; Bonilla et al. 2004b; Reiner et al. 2005; Zhu et al. 2005). The high levels of association observed over long genetic distances, lead to the expectation that phenotypes exhibiting differences between parental populations, due to some genetic factors, will also show associations with linked markers. Several independent groups have reported observing strong allelic associations between
unlinked markers spaced at substantial distances, in admixed populations (Parra et al. 1998; Lautenberger et al. 2000; McKeigue et al. 2000; Wilson and Goldstein 2000; Parra et al. 2001; Pfaff et al. 2001; Smith et al. 2001). Interest in admixture mapping has increased in recent years leading to the development of new resources (markers and methods) for this purpose (McKeigue 2000; McKeigue et al. 2000; Smith et al. 2001; Collins-Schramm et al. 2002; Hoggart et al. 2003; Hoggart et al. 2004; Montana and Pritchard 2004; Patterson et al. 2004; Smith et al. 2004).

It has been proposed that admixture mapping is a powerful technique that can facilitate identification of genes, even those with moderate effects. Complex diseases and phenotypes that have not been as amenable to linkage methods as Mendelian diseases have been may be suitable for admixture mapping. This method has been used to map genes that influence skin pigmentation (Shriver et al. 2003). Preliminary studies have also shown locus specific associations with BMI (Fernandez et al. 2003) and recently, hypertension (Zhu et al. 2005). In addition, it has been shown that ancestry levels are associated with skin pigmentation in 2 populations of African descent (Shriver et al. 2003) and several Latino populations (Bonilla et al. 2004a; Bonilla et al. 2004b), bone mineral density (Bonilla et al., 2004a), insulin sensitivity, acute insulin resistance and fasting insulin (Gower et al. 2003).

African Americans are a diverse metapopulation in whom hypertension and obesity (especially in women) is more prevalent in comparison to Americans of primarily European descent (Hodge and Zimmet 1994; Gaines and Burke 1995; Douglas et al. 1996; Allison et al. 1997). While underlying genetics may partly explain such differences in disease risk among populations, socioeconomic and psychosomatic factors may also be responsible in part of the outcome. Indeed, identifying causal mechanisms will require extensive investigation of several components, including genetics and socioeconomic factors and perhaps most importantly in characterizing the interaction between genetic and environmental factors that lead to the phenotype. This in itself is a daunting task. Instead, it may be possible to isolate parts of the causal factors to look for partial effects on the phenotype. For instance, investigating the relationship between proportional ancestry and phenotype may be one way to distinguish between the genetic and
sociocultural variables that affect complex outcomes. In this study we have used a panel of 34 ancestry informative markers (AIMs) to infer individual genomic ancestry in two African-American population samples, one from Philadelphia, PA and one from Bogalusa, LA consisting of a total of 376 individuals. Several phenotypes were characterized in these individuals including systolic and diastolic blood pressure, BMI, HDLc and Triglyceride levels. We have tested for admixture stratification and investigated the effects of ancestry on these phenotypes. In addition we have tested for specific genotype-phenotype associations. This study will help in better understanding the effects of genes that underlie complex phenotypes like hypertension and obesity.

**Materials and Methods**

**Population samples**

This study includes 376 unrelated African Americans from two locations within US, 176 individuals from Bogalusa, Louisiana and 192 from Philadelphia, Pennsylvania. The samples from Bogalusa were all collected as part of the Bogalusa Heart Study and are of mixed African, European and Indigenous American descent. This is a large community based study established in 1973 and is currently ongoing. Details regarding sample collection are available elsewhere (Voors et al. 1979; Voors et al. 1980). The purpose of this study was to investigate the early natural history of cardiovascular disease in a community consisting of European American and African Americans.

The sample from Philadelphia was collected as part of a study of blood pressure, insulin sensitivity and cardiovascular risk factors at Thomas Jefferson University. All individuals were volunteers in the study and details regarding enrollment criteria and subject characterization is described elsewhere (Cheng et al. 2004).
Phenotypic data collected on each subject includes age, sex, waist skin-fold thickness, fasting triglyceride and HDLc blood concentrations, BMI, systolic blood pressure (SBP) and diastolic blood pressure (DBP) measure. Non-normally distributed variables including triglyceride and HDLc levels were log transformed to better approximate a normal distribution. Inverse of BMI was used following previous recommendations (Flegal 1997). Non-fasting blood results were excluded. Data was also collected on years of education and used as proxy variables for socioeconomic status.

Genotyping

34 SNP ancestry informative markers (AIMs) were typed in the combined sample. These markers have been previously described and used for admixture studies (Shriver et al. 2003; Bonilla et al. 2004a; Bonilla et al. 2004b). Ancestral allele frequencies were estimated from samples of 242 Indigenous Americans (Maya, Southwestern US Indigenous Americans: Cheyenne, Pima and Pueblo), 241 Europeans (British, Irish, German and Spanish), 545 Africans (from Sierra Leone, Nigeria, Central African Republic, Ghana, Zambia, Congo, and Uganda). The program STRUCTURE 2.1 (Pritchard et al. 2000; Falush et al. 2003) was used to identify and remove individuals showing recent history of admixture. Allele frequencies, $\delta$ values and $f$ values for all three pair wise comparisons are shown in Table 3.1. All African American samples were genotyped following standard protocols (Shriver et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004).
Table 3.1: Locus ID, polymorphism, chromosomal position, allele frequency in different populations, allele frequency differences, tests of Hardy Weinberg equilibrium

<table>
<thead>
<tr>
<th>Locus</th>
<th>Alleles</th>
<th>Chromo</th>
<th>WAF</th>
<th>EU</th>
<th>IA</th>
<th>IA-EU δ</th>
<th>IA-WAF δ</th>
<th>WAF-EU δ</th>
<th>African American</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs140864</td>
<td>I/D</td>
<td>1 p34.3</td>
<td>0.124</td>
<td>0.004</td>
<td>0.584</td>
<td>0.58</td>
<td>0.46</td>
<td>0.12</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>rs16626</td>
<td>I/D</td>
<td>1 p32</td>
<td>0.759</td>
<td>0.388</td>
<td>0.318</td>
<td>0.09</td>
<td>0.46</td>
<td>0.69</td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td>rs2814778</td>
<td>T/C</td>
<td>1 q23.2</td>
<td>0.001</td>
<td>0.998</td>
<td>0.992</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.16</td>
<td>0.004</td>
</tr>
<tr>
<td>rs6003</td>
<td>A/G</td>
<td>1 q31.3</td>
<td>0.704</td>
<td>0.063</td>
<td>0.017</td>
<td>0.05</td>
<td>0.69</td>
<td>0.64</td>
<td>0.609</td>
<td>0.91</td>
</tr>
<tr>
<td>rs2065160</td>
<td>C/T</td>
<td>1 q32.1</td>
<td>0.487</td>
<td>0.917</td>
<td>0.137</td>
<td>0.78</td>
<td>0.35</td>
<td>0.43</td>
<td>0.573</td>
<td>0.03</td>
</tr>
<tr>
<td>rs2752</td>
<td>G/T</td>
<td>1 q42.2</td>
<td>0.878</td>
<td>0.433</td>
<td>0.622</td>
<td>0.19</td>
<td>0.25</td>
<td>0.44</td>
<td>0.771</td>
<td>0.52</td>
</tr>
<tr>
<td>rs3287</td>
<td>G/A</td>
<td>2 p16.1</td>
<td>0.751</td>
<td>0.215</td>
<td>0.203</td>
<td>0.03</td>
<td>0.57</td>
<td>0.54</td>
<td>0.657</td>
<td>0.43</td>
</tr>
<tr>
<td>rs17203</td>
<td>G/C</td>
<td>3 p12.1</td>
<td>0.785</td>
<td>0.133</td>
<td>0.761</td>
<td>0.69</td>
<td>0.03</td>
<td>0.65</td>
<td>0.738</td>
<td>0.69</td>
</tr>
<tr>
<td>rs7041</td>
<td>T/G</td>
<td>4 q13.3</td>
<td>0.853</td>
<td>0.156</td>
<td>0.323</td>
<td>0.18</td>
<td>0.51</td>
<td>0.70</td>
<td>0.789</td>
<td>0.1</td>
</tr>
<tr>
<td>rs4588</td>
<td>C/A</td>
<td>4 q13.3</td>
<td>0.069</td>
<td>0.607</td>
<td>0.547</td>
<td>0.07</td>
<td>0.47</td>
<td>0.54</td>
<td>0.142</td>
<td>0.34</td>
</tr>
<tr>
<td>rs16344</td>
<td>I/D</td>
<td>4q24</td>
<td>0.263</td>
<td>0.076</td>
<td>0.763</td>
<td>0.69</td>
<td>0.50</td>
<td>0.19</td>
<td>0.181</td>
<td>0.81</td>
</tr>
<tr>
<td>rs3309</td>
<td>T/A</td>
<td>5 q11.2</td>
<td>0.401</td>
<td>0.255</td>
<td>0.682</td>
<td>0.44</td>
<td>0.30</td>
<td>0.15</td>
<td>0.405</td>
<td>0.91</td>
</tr>
<tr>
<td>rs3317</td>
<td>A/G</td>
<td>5 q22.1</td>
<td>0.054</td>
<td>0.511</td>
<td>0.729</td>
<td>0.24</td>
<td>0.70</td>
<td>0.46</td>
<td>0.131</td>
<td>0.37</td>
</tr>
<tr>
<td>rs3340</td>
<td>G/A</td>
<td>5 q33.1</td>
<td>0.054</td>
<td>0.175</td>
<td>0.695</td>
<td>0.52</td>
<td>0.64</td>
<td>0.12</td>
<td>0.062</td>
<td>1</td>
</tr>
<tr>
<td>rs2763</td>
<td>C/G</td>
<td>7 p22.3</td>
<td>0.129</td>
<td>0.147</td>
<td>0.517</td>
<td>0.40</td>
<td>0.42</td>
<td>0.02</td>
<td>0.151</td>
<td>0.41</td>
</tr>
<tr>
<td>rs2161</td>
<td>A/G</td>
<td>7 q21.3</td>
<td>0.43</td>
<td>0.306</td>
<td>0.603</td>
<td>0.31</td>
<td>0.19</td>
<td>0.12</td>
<td>0.435</td>
<td>1</td>
</tr>
<tr>
<td>rs2740574</td>
<td>A/G</td>
<td>7 q22.1</td>
<td>0.198</td>
<td>0.958</td>
<td>0.953</td>
<td>0.005</td>
<td>0.53</td>
<td>0.48</td>
<td>0.324</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Table 3.1 continued

<table>
<thead>
<tr>
<th>Locus</th>
<th>Alleles</th>
<th>Chromosome</th>
<th>WAF</th>
<th>EU</th>
<th>IA</th>
<th>IA-EU δ</th>
<th>IA-WAF δ</th>
<th>WAF-EU δ</th>
<th>African American</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2740574</td>
<td>A/G</td>
<td>7 q22.1</td>
<td>0.198</td>
<td>0.958</td>
<td>0.953</td>
<td>0.005</td>
<td>0.53</td>
<td>0.48</td>
<td>0.324</td>
<td>0.14</td>
</tr>
<tr>
<td>rs285</td>
<td>T/C</td>
<td>8 p21.3</td>
<td>0.971</td>
<td>0.492</td>
<td>0.45</td>
<td>0.70</td>
<td>0.62</td>
<td>0.08</td>
<td>0.884</td>
<td>0.04</td>
</tr>
<tr>
<td>rs2695</td>
<td>G/A</td>
<td>9 q21.31</td>
<td>0.805</td>
<td>0.881</td>
<td>0.218</td>
<td>0.40</td>
<td>0.03</td>
<td>0.43</td>
<td>0.766</td>
<td>0.55</td>
</tr>
<tr>
<td>rs594689</td>
<td>T/C</td>
<td>11 q13.3</td>
<td>0.087</td>
<td>0.516</td>
<td>0.14</td>
<td>0.41</td>
<td>0.03</td>
<td>0.44</td>
<td>0.135</td>
<td>0.82</td>
</tr>
<tr>
<td>rs1042602</td>
<td>A/C</td>
<td>11 q14.3</td>
<td>0.005</td>
<td>0.449</td>
<td>0.06</td>
<td>0.63</td>
<td>0.09</td>
<td>0.54</td>
<td>0.06</td>
<td>1</td>
</tr>
<tr>
<td>rs1800498</td>
<td>T/C</td>
<td>11 q23.2</td>
<td>0.118</td>
<td>0.67</td>
<td>0.089</td>
<td>0.52</td>
<td>0.60</td>
<td>0.08</td>
<td>0.187</td>
<td>0.16</td>
</tr>
<tr>
<td>rs1079598</td>
<td>C/T</td>
<td>11 q23.2</td>
<td>0.063</td>
<td>0.144</td>
<td>0.628</td>
<td>0.03</td>
<td>0.43</td>
<td>0.46</td>
<td>0.084</td>
<td>1</td>
</tr>
<tr>
<td>rs5443</td>
<td>T/C</td>
<td>12 p13.31</td>
<td>0.795</td>
<td>0.332</td>
<td>0.365</td>
<td>0.14</td>
<td>0.75</td>
<td>0.61</td>
<td>0.777</td>
<td>0.44</td>
</tr>
<tr>
<td>rs2252544</td>
<td>G/A</td>
<td>13 q14.2</td>
<td>0.926</td>
<td>0.315</td>
<td>0.187</td>
<td>0.26</td>
<td>0.37</td>
<td>0.63</td>
<td>0.853</td>
<td>1</td>
</tr>
<tr>
<td>rs1800404</td>
<td>A/G</td>
<td>15 q12</td>
<td>0.115</td>
<td>0.746</td>
<td>0.484</td>
<td>0.51</td>
<td>0.33</td>
<td>0.18</td>
<td>0.191</td>
<td>0.86</td>
</tr>
<tr>
<td>rs2862</td>
<td>C/T</td>
<td>15 q14</td>
<td>0.386</td>
<td>0.201</td>
<td>0.696</td>
<td>0.45</td>
<td>0.41</td>
<td>0.04</td>
<td>0.338</td>
<td>0.91</td>
</tr>
<tr>
<td>rs4646</td>
<td>T/G</td>
<td>15 q21.2</td>
<td>0.34</td>
<td>0.296</td>
<td>0.719</td>
<td>0.13</td>
<td>0.48</td>
<td>0.35</td>
<td>0.255</td>
<td>0.007</td>
</tr>
<tr>
<td>rs2228478</td>
<td>G/A</td>
<td>16 q24.3</td>
<td>0.59</td>
<td>0.163</td>
<td>0.072</td>
<td>0.05</td>
<td>0.39</td>
<td>0.45</td>
<td>0.475</td>
<td>0.16</td>
</tr>
<tr>
<td>rs2891</td>
<td>C/T</td>
<td>17p13</td>
<td>0.04</td>
<td>0.47</td>
<td>0.42</td>
<td>0.42</td>
<td>0.06</td>
<td>0.48</td>
<td>0.094</td>
<td>0.54</td>
</tr>
<tr>
<td>rs2816</td>
<td>C/T</td>
<td>17p13.1</td>
<td>0</td>
<td>0.48</td>
<td>0.058</td>
<td>0.59</td>
<td>0.74</td>
<td>0.15</td>
<td>0.05</td>
<td>0.58</td>
</tr>
<tr>
<td>rs4884</td>
<td>A/G</td>
<td>19q13.2</td>
<td>0.164</td>
<td>0.314</td>
<td>0.903</td>
<td>0.06</td>
<td>0.39</td>
<td>0.44</td>
<td>0.190</td>
<td>0.86</td>
</tr>
<tr>
<td>rs3138520</td>
<td>A/G</td>
<td>19q13.3</td>
<td>0.81</td>
<td>0.36</td>
<td>0.419</td>
<td>0.73</td>
<td>0.18</td>
<td>0.55</td>
<td>0.746</td>
<td>0.57</td>
</tr>
<tr>
<td>rs6058017</td>
<td>A/G</td>
<td>20q11</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.298</td>
<td>0.86</td>
</tr>
<tr>
<td>rs16383</td>
<td>D/I</td>
<td>22q13.2</td>
<td>0.261</td>
<td>0.81</td>
<td>0.081</td>
<td>0.05</td>
<td>0.53</td>
<td>0.48</td>
<td>0.364</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Data Analysis

Individual ancestry estimates were calculated using a maximum likelihood method as described by Chakraborty, 1986 and implemented in the program MLIAE, written for this purpose and provided in Appendix E. Allele frequency estimates and exact tests for HWE were performed using Genepop (Rousset and Raymond 1995) software separately for both populations and for the combined sample. Two separate tests for detecting admixture stratification in the sample were conducted. In one test we compared the proportion of unlinked marker pairs with the expected value at the 5% significance level. Gametic disequilibrium was estimated using the program 3Locus (Long et al. 1995), which implements an expectation maximization algorithm and also calculates a likelihood ratio statistic (G) to ascertain allelic associations. In a second test for structure we randomly split the marker panels into two non-syntenic subsets and ascertained ancestry proportions independently with each subset. This procedure was repeated twenty times, by randomly selecting the marker subsets. We then tested for correlation in ancestry estimates obtained with the two different subsets in each run.

We used linear regression models available in SPSS v.10 to test the effect of ancestry on each phenotype as a dependent variable. BMI, SBP, DBP, HDLc, triglyceride were all used as continuous variables. In addition we used two BP measures to create a dichotomous variable for hypertension status, where individuals with DBP > 85 and/or SBP > 135 were classified as hypertensive. Triglycerides and HDLc values were log transformed and inverse of BMI was used in all analyses to improve normality and all subsequent references to these variables refer to the transformed variables. Impact of SES on each of the outcome variables was specifically tested. Age, sex, SES and BMI were included as covariates for all other phenotypes. Age, sex and SES were used as covariates for BMI. In a second set of linear regression models city of origin (location) was included as a covariate along with other covariates mentioned previously. We next tested for the association of each outcome variable as the dependent variable with each of the marker
genotypes as independent variable in analysis of variance (ANOVA) models. These analyses were repeated by controlling for individual ancestry proportions (ANOVA/IAE). This method has previously been used to map genes underlying variation in skin pigmentation (Shriver et al. 2003; Bonilla et al. 2004a; Bonilla et al. 2004b), and BMI, BMD and total fat mass (Bonilla et al. 2004a). Bonferroni correction was applied for all tests and significance levels were accordingly modified to account for effects of multiple testing.

In a complimentary approach we have used Bayesian MCMC methods implemented in the program ADMIXMAP (McKeigue 2000; McKeigue et al. 2000; Hoggart et al. 2003; Bonilla et al. 2004a; Hoggart et al. 2004) to ascertain individual ancestry proportions, test for admixture stratification and test for effects of ancestry on phenotypes. ADMIXMAP is designed for analyzing datasets that consist of trait measurements and genotype data on a sample of individuals from an admixed population. The program requires ancestry informative markers, which have large allele frequency differences between ancestral populations to efficiently model admixture. ADMIXMAP uses a combination of Bayesian and classical theory. A probability model is specified with (non-informative) prior values for distribution of admixture in the population, and the posterior distribution of all unobserved variables is generated using Markov chain Monte Carlo (MCMC) simulation, conditional on observed genotypes and trait values. Unobserved values modeled include ancestral state (sub-population of origin) on each gamete at each locus, and the ancestry-specific allele frequencies at each locus. ADMIXMAP can also test to test if the population has been specified with the correct number of sub-populations (McKeigue 2000; McKeigue et al. 2000; Hoggart et al. 2003; Bonilla et al. 2004; Hoggart et al. 2004) to account for all residual stratification in the sample and is thus a test for stratification.

Several score tests based on likelihood of missing data are built into the program to test for linkage between genotype and phenotypes. The program fits a model based on the null hypothesis and this null hypothesis can be tested against alternatives with a score test computed by averaging over the posterior distribution of the missing data. The model to be tested is based on a regression equation using a vector of observed alleles or
haplotypes at a locus and compared against a null hypothesis of no association. The program can be used to test for locus specific effects on a phenotype as well as effects of ancestry at each locus on the phenotype. Further details regarding theory and application are given elsewhere (McKeigue et al. 2000; Hoggart et al. 2003; Hoggart et al. 2004).

**Results**

**Sample characteristics**

Average estimates of phenotypic measures for the combined sample are shown in Table 3.2. Average measures of phenotypes for both study sites are shown separately in Table 3.3. Mean age was significantly higher in the individuals from Philadelphia compared to those from Bogalusa. Height and weight did not differ significantly between locations. Mean levels of SBP and DBP are significantly higher (P<0.0001) in the Philadelphia sample. Mean levels of HDLc are significantly higher in the Bogalusa sample. Triglyceride and BMI levels do not differ significantly between locations.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Mean ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>30.91 ± 5.6</td>
</tr>
<tr>
<td>Height</td>
<td>168.4 ± 9.4</td>
</tr>
<tr>
<td>Weight</td>
<td>86.48 ± 23.5</td>
</tr>
<tr>
<td>SBP</td>
<td>120.36 ± 16.2</td>
</tr>
<tr>
<td>DBP</td>
<td>72.25 ± 13.8</td>
</tr>
<tr>
<td>BMI</td>
<td>30.51 ± 8</td>
</tr>
<tr>
<td>HDLc</td>
<td>50.5 ± 15.7</td>
</tr>
<tr>
<td>Triglyceride</td>
<td>93.52 ± 56.6</td>
</tr>
</tbody>
</table>
Table 3.3: Characteristics of sample from each location

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Bogalusa (N=174)</th>
<th>Philadelphia(N=202)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age*</td>
<td>29.2 ± 5.5</td>
<td>32.5 ± 5.3</td>
</tr>
<tr>
<td>Height</td>
<td>168.4 ± 9.1</td>
<td>168.5 ± 9.7</td>
</tr>
<tr>
<td>Weight</td>
<td>84.8 ± 25</td>
<td>87.9 ± 22</td>
</tr>
<tr>
<td>SBP*</td>
<td>114.6 ± 13.4</td>
<td>125.3 ± 26.76</td>
</tr>
<tr>
<td>DBP*</td>
<td>69.1 ± 10.9</td>
<td>80.6 ± 13.9</td>
</tr>
<tr>
<td>BMI</td>
<td>29.8 ± 8.1</td>
<td>31.1 ± 8</td>
</tr>
<tr>
<td>HDLc*</td>
<td>52.5 ± 15.5</td>
<td>48.7 ± 15.72</td>
</tr>
<tr>
<td>Triglyceride</td>
<td>96.2 ± 63.1</td>
<td>91.2 ± 50.22</td>
</tr>
</tbody>
</table>

* Significantly different between groups (P<0.01)

Sex specific estimates for phenotype in the total sample are shown in Table 3.4 and for each study site separately are shown in Table 3.5. In the combined sample males were taller and weighed more than the females, in addition males had significantly higher SBP and triglyceride levels. Females had significantly higher BMI and HDLc. All subsequent phenotype analyses were performed on the combined sample to increase power in the study.

Table 3.4: Sex specific estimates of phenotypes in total sample

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Males (N=134)</th>
<th>Females (N=242)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>30.3 ± 5</td>
<td>31.1 ± 6</td>
</tr>
<tr>
<td>Height*</td>
<td>177.03 ± 6.7</td>
<td>163.7 ± 7</td>
</tr>
<tr>
<td>Weight*</td>
<td>90.8 ± 23.6</td>
<td>84.4 ± 23.1</td>
</tr>
<tr>
<td>SBP*</td>
<td>123.6 ± 13.5</td>
<td>118.5 ± 17.36</td>
</tr>
<tr>
<td>DBP</td>
<td>76.51 ± 12.65</td>
<td>80.6 ± 13.9</td>
</tr>
<tr>
<td>BMI*</td>
<td>29 ± 7.6</td>
<td>31.5 ± 8.3</td>
</tr>
<tr>
<td>HDLc*</td>
<td>49.19 ± 19.4</td>
<td>51.12 ± 13.3</td>
</tr>
<tr>
<td>Triglyceride*</td>
<td>111.27 ± 71.65</td>
<td>83.99 ± 43.73</td>
</tr>
</tbody>
</table>

* Significantly different between sex (P<0.05)
Table 3.5: Sex specific distributions of phenotypes in different locations

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Bogalusa Males (N=62)</th>
<th>Bogalusa Females (N=112)</th>
<th>Philadelphia Males (N=72)</th>
<th>Philadelphia Females (N=130)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>29.5 ± 5.5</td>
<td>28.8 ± 5.4</td>
<td>31.3 ± 4.4*</td>
<td>33.2 ± 5.5</td>
</tr>
<tr>
<td>Height</td>
<td>176.7 ± 6*</td>
<td>163.9 ± 7.03</td>
<td>177.4 ± 7.2*</td>
<td>163.5 ± 6.8</td>
</tr>
<tr>
<td>Weight</td>
<td>91 ± 26.6*</td>
<td>81.4 ± 23.8</td>
<td>90.8 ± 21.2</td>
<td>86.3 ± 22.3</td>
</tr>
<tr>
<td>SBP</td>
<td>118.8 ± 11.5*</td>
<td>112.3 ± 13.36</td>
<td>127.8 ± 13.5</td>
<td>124 ± 18.2</td>
</tr>
<tr>
<td>DBP</td>
<td>70.56 ± 9.57</td>
<td>68.24 ± 11.52</td>
<td>81.3 ± 12.8</td>
<td>80.17 ± 14.54</td>
</tr>
<tr>
<td>BMI</td>
<td>29.15 ± 8.2</td>
<td>30.19 ± 8</td>
<td>28.9 ± 6.5*</td>
<td>32.3 ± 8.5</td>
</tr>
<tr>
<td>HDLc</td>
<td>49.08 ± 16.4*</td>
<td>54.4 ± 14.6</td>
<td>49.3 ± 21.7</td>
<td>48.3 ± 11.2</td>
</tr>
<tr>
<td>Triglyceride</td>
<td>117.9 ± 83.7*</td>
<td>84.1 ± 44.1</td>
<td>105.3 ± 58.7*</td>
<td>83.4 ± 43.2</td>
</tr>
</tbody>
</table>

* Significantly different between sexes within each city

Proportional ancestry and admixture stratification

We estimated ancestry in this sample assuming a three way admixture model between Europeans, West Africans (WAF) and Indigenous Americans (IA). Proportional ML estimates of ancestral contributions to the combined sample are 0.84 ± 0.14 WAF, 0.12 ± 0.12 EU and 0.05 ± 0.08 IA. Individually, mean estimates in the Philadelphia sample are 0.84 ± 0.14 WAF, 0.12 ± 0.13 EU and 0.04 ± 0.07 IA. Mean estimates in the Bogalusa sample are 0.84 ± 0.14 WAF, 0.12 ± 0.12 EU and 0.05 ± 0.08 IA. It should be pointed that these are means of individual ancestry estimates for the sample and not group estimates. Figure 3.1 shows the distribution of individual genomic ancestry in the two samples. We found neither significant difference in mean individual ancestry proportions between the two locations using either t-test nor differences in distributions of proportional ancestry using the Kolmogorov-Smirnov test. We did not observe sex specific differences in ancestry in the total sample or in each study site separately (data not shown). Figure 3.2 shows the histogram of distribution of individual genomic
ancestry estimates in the combined sample. The IAC test shows strong correlation between estimates obtained with different subsets of markers ($R^2 = 0.89$, $P<0.0001$ for WAF, $R^2 = 0.65$, $P<0.0001$ for EU and $R^2 = 0.32$, $P=0.005$ for IA) indicating presence of stratification in the total sample. We observed significant non-random association between 6% (31 out of 512) of unlinked marker pairs in the sample. At the 5% significance level we expect 25 of 512 unlinked markers to show an association by chance. In addition we observed significant association between the markers rs2814778 and rs6003 ($P = 0.04$) which are 40 cM apart, markers rs594689 and rs1042602 ($P = 0.04$) which are 21 cM apart, rs1042602 and rs1800498 ($P = 0.008$) which are 21 cM apart and between markers rs2891 and rs2816 ($P <0.005$) which are 10 cM apart.

Figure 3.1: Distribution of ancestry estimates in African American samples from Bogalusa, LA (N=176; dots) and Philadelphia, PA (N=202, blank circles).
Figure 3.2: Histogram of distribution of ancestry proportions in an African-American metapopulation. Proportional ancestry is shown on X axis and number of individuals on the Y axis.
**Ancestry component – phenotype associations**

Using linear regression models we observed no significant association between SES and BMI or between SES and triglyceride levels. SES shows suggestive correlation with HDLc ($R^2 = 1.1\%$, $P = 0.048$) and significant correlation with both SBP ($R^2 = 9.3\%$, $P<0.0001$), and DBP ($R^2 = 16.8\%$, $P<0.0001$). All phenotypes were adjusted for age, sex, location and SES in subsequent analyses. In addition, SBP, DBP, triglycerides and HDLc measures were also adjusted for BMI.

BMI shows suggestive association with proportional IA ancestry ($R^2 = 4.6\%$ and $P = 0.026$). No associations were detected between proportional WAF ancestry or with proportional EU ancestry. Multiple regression models, which include age, sex, location, SES, and each of the ancestry axes separately were significant (WAF: $P = 0.004$, IA: $P = 0.001$, EU: $P = 0.011$). A negative association was seen with European and Indigenous American ancestry and a positive association was seen with West African ancestry in the multiple regression models. In each case $R^2$ increased by the addition of ancestry estimates as a predictor variable, which indicates that including ancestry in the model helps explain additional variation in the phenotype. However, when BMI was coded as a binary variable (BMI>30 labeled as Obese) only the IA ancestry showed a suggestive negative association ($R^2 = 2.9\%$, $P = 0.03$).

HDLc levels did not show any association with ancestry (WAF: $P = 0.964$, EU: $P = 0.868$). Neither did similarly adjusted levels of triglycerides (WAF: $P = 0.326$, EU: $P = 0.688$). In all multiple regression models for HDLc and Triglycerides, addition of ancestry estimates did not increase $R^2$ values any further. All results were the same when analyses were repeated by leaving out location from the model.

A multiple regression model for SBP as an outcome including age, sex, location, SES and BMI as explanatory variables explains $27.2\%$ ($P<0.005$) of variation in the phenotype. Neither WAF ($P = 0.13$) nor IA ($P = 0.556$) ancestry is significantly
associated with adjusted SBP values, EU ancestry shows suggestive negative association with adjusted SBP ($R^2 = 28\%$ and $P = 0.047$). For DBP as an outcome, $30\%$ ($P<0.005$) of variation in the phenotype is accounted for in the multiple regression model. Neither WAF ($P = 0.193$) nor IA (0.356) is significantly associated with adjusted DBP. Only EU shows a suggestive negative association with adjusted DBP ($R^2 = 30.8, P = 0.049$).

Since blood pressure and BMI are more prevalent in African-American females than in males, all analyses were repeated in the two sexes separately. No associations were observed in the males. In females, overall trends were similar as observed in the total sample. Adjusted BMI, shows suggestive association with proportional IA ancestry ($R^2 = 1.6\%, P = 0.03$). Adjusted SBP and DBP both show suggestive negative association with proportional EU ancestry ($R^2 = 2.1\%, P = 0.016$ for SBP and $R^2 = 1.3\%, P = 0.047$ for DBP).

**Locus specific associations**

In the next set of tests we investigated genotype-phenotype associations with and without controlling for ancestry using ANOVA. Bonferroni correction was applied to correct for multiple tests. The results for all phenotypes are shown in Table 3.6. For BMI as a continuous variable and separately using BMI $>30$ to define obesity status and using this value as a binary response variable, we observed strong associations between BMI and locus rs3317 which persisted even after correcting for ancestry. No locus specific associations were observed in triglycerides or HDLc. No associations were seen for BP measures as a continuous variable, prior to correcting for proportional ancestry. After correcting for individual ancestry, locus rs6003 shows a significant association with SBP. When hypertension was coded as a binary response variable, the locus rs16626 shows a significant association that persists after controlling for ancestry. Table 3.6 also shows the results from the ADMIXMAP tests that are discussed later.
Table 3.6: Associations between genotypes and phenotypes

<table>
<thead>
<tr>
<th>Locus</th>
<th>BMI</th>
<th>OBESITY</th>
<th>HDLc</th>
<th>TRIGLYCERIDES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>A/IAE</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>rs140864</td>
<td>0.975</td>
<td>0.823</td>
<td>0.99</td>
<td>0.808</td>
</tr>
<tr>
<td>rs16626</td>
<td>0.124</td>
<td>0.127</td>
<td>0.13</td>
<td>0.691</td>
</tr>
<tr>
<td>rs2814778</td>
<td>0.406</td>
<td>0.942</td>
<td>0.5</td>
<td>0.75</td>
</tr>
<tr>
<td>rs6003</td>
<td>0.928</td>
<td>0.681</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td>rs2065160</td>
<td>0.547</td>
<td>0.583</td>
<td>0.56</td>
<td>0.053</td>
</tr>
<tr>
<td>rs2752</td>
<td>0.537</td>
<td>0.587</td>
<td>0.51</td>
<td>0.563</td>
</tr>
<tr>
<td>rs3287</td>
<td>0.54</td>
<td>0.32</td>
<td>0.51</td>
<td>0.641</td>
</tr>
<tr>
<td>rs17203</td>
<td>0.955</td>
<td>0.895</td>
<td>0.83</td>
<td>0.881</td>
</tr>
<tr>
<td>rs7041</td>
<td>0.785</td>
<td>0.985</td>
<td>0.54</td>
<td>0.346</td>
</tr>
<tr>
<td>rs163444</td>
<td>0.777</td>
<td>0.022</td>
<td>0.74</td>
<td>0.976</td>
</tr>
<tr>
<td>rs3309</td>
<td>0.173</td>
<td>0.221</td>
<td>0.18</td>
<td>0.131</td>
</tr>
<tr>
<td>rs3317</td>
<td>&gt;0.0005</td>
<td>&gt;0.0005</td>
<td>0.00018</td>
<td>&gt;0.0005</td>
</tr>
<tr>
<td>rs3340</td>
<td>0.36</td>
<td>0.57</td>
<td>0.35</td>
<td>0.824</td>
</tr>
<tr>
<td>rs2763</td>
<td>0.822</td>
<td>0.856</td>
<td>0.91</td>
<td>0.696</td>
</tr>
<tr>
<td>rs2161</td>
<td>0.69</td>
<td>0.698</td>
<td>0.97</td>
<td>0.911</td>
</tr>
<tr>
<td>rs2740574</td>
<td>0.323</td>
<td>0.656</td>
<td>0.26</td>
<td>0.521</td>
</tr>
<tr>
<td>rs285</td>
<td>0.664</td>
<td>0.961</td>
<td>0.79</td>
<td>0.471</td>
</tr>
<tr>
<td>rs2695</td>
<td>0.093</td>
<td>0.14</td>
<td>0.086</td>
<td>0.268</td>
</tr>
<tr>
<td>rs594689</td>
<td>0.28</td>
<td>0.456</td>
<td>0.29</td>
<td>0.897</td>
</tr>
</tbody>
</table>

A: ANOVA; A/IAE: ANOVA controlling for individual estimates of ancestry; P: Bayesian P values from ADMIXMAP. All significant associations (after applying Bonferroni correction) are shown in bold.
Table 3.6 continued.

<table>
<thead>
<tr>
<th>Locus</th>
<th>A</th>
<th>A/IAE</th>
<th>p</th>
<th>A</th>
<th>A/IAE</th>
<th>p</th>
<th>A</th>
<th>A/IAE</th>
<th>p</th>
<th>A</th>
<th>A/IAE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1042602</td>
<td>0.705</td>
<td>0.966</td>
<td>0.66</td>
<td>0.81</td>
<td>0.603</td>
<td>0.83</td>
<td>0.676</td>
<td>0.833</td>
<td>0.98</td>
<td>0.453</td>
<td>0.253</td>
<td>0.55</td>
</tr>
<tr>
<td>rs1800498</td>
<td>0.549</td>
<td>0.083</td>
<td>0.44</td>
<td>0.578</td>
<td>0.387</td>
<td>0.49</td>
<td>0.189</td>
<td>0.387</td>
<td>0.26</td>
<td>0.694</td>
<td>0.502</td>
<td>0.49</td>
</tr>
<tr>
<td>rs1079598</td>
<td>0.242</td>
<td>0.235</td>
<td>0.27</td>
<td>0.201</td>
<td>0.204</td>
<td>0.22</td>
<td>0.093</td>
<td>0.072</td>
<td>0.97</td>
<td>0.225</td>
<td>0.133</td>
<td>0.39</td>
</tr>
<tr>
<td>rs5443</td>
<td>0.493</td>
<td>0.171</td>
<td>0.51</td>
<td>0.486</td>
<td>0.215</td>
<td>0.45</td>
<td>0.683</td>
<td>0.845</td>
<td>0.61</td>
<td>0.819</td>
<td>0.822</td>
<td>0.83</td>
</tr>
<tr>
<td>rs2252544</td>
<td>0.373</td>
<td>0.38</td>
<td>0.31</td>
<td>0.283</td>
<td>0.28</td>
<td>0.22</td>
<td>0.866</td>
<td>0.754</td>
<td>0.81</td>
<td>0.903</td>
<td>0.937</td>
<td>0.66</td>
</tr>
<tr>
<td>rs1800404</td>
<td>0.883</td>
<td>0.739</td>
<td>0.87</td>
<td>0.356</td>
<td>0.522</td>
<td>0.33</td>
<td>0.356</td>
<td>0.429</td>
<td>0.51</td>
<td>0.282</td>
<td>0.402</td>
<td>0.27</td>
</tr>
<tr>
<td>rs2862</td>
<td>0.739</td>
<td>0.659</td>
<td>0.73</td>
<td>0.822</td>
<td>0.747</td>
<td>0.81</td>
<td>0.625</td>
<td>0.654</td>
<td>0.71</td>
<td>0.991</td>
<td>0.915</td>
<td>0.91</td>
</tr>
<tr>
<td>rs4646</td>
<td>0.228</td>
<td>0.269</td>
<td>0.25</td>
<td>0.281</td>
<td>0.318</td>
<td>0.29</td>
<td>0.742</td>
<td>0.771</td>
<td>0.57</td>
<td>0.378</td>
<td>0.404</td>
<td>0.26</td>
</tr>
<tr>
<td>rs2228478</td>
<td>0.213</td>
<td>0.297</td>
<td>0.27</td>
<td>0.503</td>
<td>0.606</td>
<td>0.56</td>
<td>0.897</td>
<td>0.815</td>
<td>0.91</td>
<td>0.098</td>
<td>0.095</td>
<td>0.31</td>
</tr>
<tr>
<td>rs2891</td>
<td>0.207</td>
<td>0.427</td>
<td>0.22</td>
<td>0.212</td>
<td>0.337</td>
<td>0.2</td>
<td>0.538</td>
<td>0.666</td>
<td>0.84</td>
<td>0.52</td>
<td>0.36</td>
<td>0.79</td>
</tr>
<tr>
<td>rs2816</td>
<td>0.305</td>
<td>0.636</td>
<td>0.31</td>
<td>0.537</td>
<td>0.828</td>
<td>0.55</td>
<td>0.491</td>
<td>0.699</td>
<td>0.62</td>
<td>0.583</td>
<td>0.35</td>
<td>0.78</td>
</tr>
<tr>
<td>rs4884</td>
<td>0.246</td>
<td>0.327</td>
<td>0.29</td>
<td>0.371</td>
<td>0.452</td>
<td>0.39</td>
<td>0.02</td>
<td>0.025</td>
<td>0.016</td>
<td>0.614</td>
<td>0.527</td>
<td>0.79</td>
</tr>
<tr>
<td>rs3138520</td>
<td>0.176</td>
<td>0.096</td>
<td>0.17</td>
<td>0.164</td>
<td>0.091</td>
<td>0.14</td>
<td>0.619</td>
<td>0.537</td>
<td>0.55</td>
<td>0.412</td>
<td>0.562</td>
<td>0.77</td>
</tr>
<tr>
<td>rs6058017</td>
<td>0.854</td>
<td>0.899</td>
<td>0.67</td>
<td>0.532</td>
<td>0.405</td>
<td>0.62</td>
<td>0.761</td>
<td>0.614</td>
<td>0.64</td>
<td>0.175</td>
<td>0.134</td>
<td>0.25</td>
</tr>
<tr>
<td>rs16383</td>
<td>0.337</td>
<td>0.442</td>
<td>0.34</td>
<td>0.246</td>
<td>0.298</td>
<td>0.23</td>
<td>0.003</td>
<td>0.005</td>
<td>0.00055</td>
<td>0.708</td>
<td>0.757</td>
<td>0.57</td>
</tr>
</tbody>
</table>

A: ANOVA; A/IAE: ANOVA controlling for individual estimates of ancestry; P: Bayesian P values from ADMIXMAP. All significant associations (after applying Bonferroni correction) are shown in bold.
<table>
<thead>
<tr>
<th>Locus</th>
<th>SBP</th>
<th>DBP</th>
<th>Hypertension</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A A/IAE P</td>
<td>A A/IAE P</td>
<td>A A/IAE P</td>
</tr>
<tr>
<td>rs140864</td>
<td>0.637 0.649 0.79</td>
<td>0.98 0.89 0.95</td>
<td>0.658 0.658 0.65</td>
</tr>
<tr>
<td>rs16626</td>
<td>0.076 0.085 0.6</td>
<td>0.214 0.205 0.5</td>
<td><strong>0.001</strong> 0.001 0.0093</td>
</tr>
<tr>
<td>rs2814778</td>
<td>0.875 0.473 0.6</td>
<td>0.179 0.058 0.79</td>
<td>0.436 0.445 0.43</td>
</tr>
<tr>
<td>rs6003</td>
<td>0.006 <strong>0.001</strong> 0.0061</td>
<td>0.179 0.065 0.18</td>
<td>0.067 0.028 0.048</td>
</tr>
<tr>
<td>rs2065160</td>
<td>0.054 0.027 0.046</td>
<td>0.226 0.209 0.28</td>
<td>0.314 0.308 0.32</td>
</tr>
<tr>
<td>rs2752</td>
<td>0.142 0.161 0.48</td>
<td>0.052 0.055 0.14</td>
<td>0.433 0.766 0.49</td>
</tr>
<tr>
<td>rs3287</td>
<td>0.589 0.402 0.76</td>
<td>0.742 0.302 0.63</td>
<td>0.532 0.735 0.52</td>
</tr>
<tr>
<td>rs17203</td>
<td>0.032 0.046 0.92</td>
<td>0.27 0.3 0.25</td>
<td>0.754 0.317 0.65</td>
</tr>
<tr>
<td>rs7041</td>
<td>0.897 0.763 0.98</td>
<td>0.564 0.984 0.96</td>
<td>0.674 0.564 0.35</td>
</tr>
<tr>
<td>rs16344</td>
<td>0.985 0.135 0.96</td>
<td>0.27 0.3 0.25</td>
<td>0.674 0.564 0.35</td>
</tr>
<tr>
<td>rs3309</td>
<td>0.339 0.394 0.68</td>
<td>0.279 0.318 0.3</td>
<td>0.38 0.404 0.44</td>
</tr>
<tr>
<td>rs3317</td>
<td>0.419 0.666 0.02</td>
<td>0.918 0.615 0.92</td>
<td>0.753 0.864 0.83</td>
</tr>
<tr>
<td>rs3340</td>
<td>0.626 0.448 0.45</td>
<td>0.683 0.52 0.55</td>
<td>0.676 0.72 0.68</td>
</tr>
<tr>
<td>rs2763</td>
<td>0.894 0.927 0.82</td>
<td>0.876 0.846 0.83</td>
<td>0.556 0.543 0.54</td>
</tr>
<tr>
<td>rs2161</td>
<td>0.949 0.934 0.67</td>
<td>0.77 0.561 0.93</td>
<td>0.842 0.861 0.74</td>
</tr>
<tr>
<td>rs2740574</td>
<td>0.108 0.179 0.47</td>
<td>0.422 0.555 0.64</td>
<td>0.272 0.845 0.28</td>
</tr>
<tr>
<td>rs285</td>
<td>0.744 0.456 0.52</td>
<td>0.937 0.158 0.94</td>
<td>0.925 0.94 0.96</td>
</tr>
<tr>
<td>rs2695</td>
<td>0.523 0.246 0.93</td>
<td>0.928 0.956 0.75</td>
<td>0.676 0.644 0.78</td>
</tr>
<tr>
<td>rs594689</td>
<td>0.51 0.847 0.84</td>
<td>0.489 0.778 0.7</td>
<td>0.392 0.461 0.37</td>
</tr>
<tr>
<td>rs1042602</td>
<td>0.008 0.02 0.013</td>
<td>0.002 0.004 <strong>0.0016</strong></td>
<td>0.103 0.165 0.078</td>
</tr>
<tr>
<td>rs1800498</td>
<td>0.764 0.469 0.78</td>
<td>0.425 0.323 0.53</td>
<td>0.747 0.473 0.65</td>
</tr>
<tr>
<td>rs1079598</td>
<td>0.937 0.84 0.9</td>
<td>0.409 0.396 0.57</td>
<td>0.886 0.902 0.9</td>
</tr>
<tr>
<td>rs5443</td>
<td>0.598 0.999 0.35</td>
<td>0.611 0.997 0.54</td>
<td>0.417 0.52 0.4</td>
</tr>
<tr>
<td>rs2252544</td>
<td>0.862 0.757 0.49</td>
<td>0.334 0.328 0.29</td>
<td>0.359 0.993 0.51</td>
</tr>
<tr>
<td>rs1800404</td>
<td>0.127 0.336 0.17</td>
<td>0.114 0.203 0.18</td>
<td>0.531 0.649 0.52</td>
</tr>
<tr>
<td>rs2862</td>
<td>0.257 0.091 0.24</td>
<td>0.847 0.768 0.89</td>
<td>0.179 0.165 0.24</td>
</tr>
<tr>
<td>rs4646</td>
<td>0.359 0.418 0.63</td>
<td>0.599 0.614 0.58</td>
<td>0.281 0.307 0.35</td>
</tr>
<tr>
<td>rs2228478</td>
<td>0.415 0.334 0.16</td>
<td>0.684 0.614 0.67</td>
<td>0.27 0.27 0.3</td>
</tr>
<tr>
<td>rs2891</td>
<td>0.048 0.095 0.17</td>
<td>0.215 0.336 0.27</td>
<td>0.606 0.655 0.57</td>
</tr>
<tr>
<td>rs2816</td>
<td>0.853 0.559 0.6</td>
<td>0.206 0.098 0.2</td>
<td>0.261 0.166 0.29</td>
</tr>
<tr>
<td>rs4884</td>
<td>0.414 0.18 0.72</td>
<td>0.466 0.281 0.49</td>
<td>0.424 0.381 0.4</td>
</tr>
<tr>
<td>rs3138520</td>
<td>0.32 0.503 0.14</td>
<td>0.544 0.746 0.45</td>
<td>0.758 0.849 0.76</td>
</tr>
<tr>
<td>rs6058017</td>
<td>0.535 0.444 0.51</td>
<td>0.613 0.522 0.73</td>
<td>0.055 0.043 0.051</td>
</tr>
<tr>
<td>rs16383</td>
<td>0.214 0.173 0.1</td>
<td>0.681 0.608 0.73</td>
<td>0.644 0.591 0.58</td>
</tr>
</tbody>
</table>

A: ANOVA; A/IAE: ANOVA controlling for individual estimates of ancestry; P: Bayesian P values from ADMIXMAP. All significant associations (after applying Bonferroni correction) are shown in bold.
**ADMIXMAP results**

As a complimentary approach, we used several tests built into ADMIXMAP to analyze this sample. The test for stratification was conducted by running the program separately for 1, 2 and 3 subpopulations. Corresponding test statistics are 0.148, 0.793 and 0.792. This indicates that using a model with two subpopulations adequately models the data and there is no residual confounding due to stratification that is unaccounted for.

Table 3.7 shows the means and 95% credible interval estimates of posterior parameters obtained using the ADMIXMAP. Mean estimates of proportional ancestry are 0.85 WAF, 0.14 EU and 0.01 IA, which is similar to estimates obtained with ML. Figure 3.3 shows the distribution of ancestry estimates for each ancestry axis. In addition, these figures show the fit of the data to the assumed model (proportional ancestry modeled using the Dirichlet distribution) and indeed, the data fits the model well.

Table 3.7: Summary of posterior parameters estimated using ADMIXMAP

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Pct2.5</th>
<th>Pct97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportional WAF</td>
<td>0.854</td>
<td>0.840</td>
<td>0.869</td>
</tr>
<tr>
<td>Proportional EU</td>
<td>0.131</td>
<td>0.117</td>
<td>0.147</td>
</tr>
<tr>
<td>Proportional IA</td>
<td>0.015</td>
<td>0.009</td>
<td>0.024</td>
</tr>
<tr>
<td>sumIntensities</td>
<td>0.715</td>
<td>0.422</td>
<td>1.020</td>
</tr>
<tr>
<td>Dirichlet.WAF</td>
<td>7.960</td>
<td>5.670</td>
<td>11.400</td>
</tr>
<tr>
<td>Dirichlet.EU</td>
<td>1.230</td>
<td>0.854</td>
<td>1.730</td>
</tr>
<tr>
<td>Dirichlet.IA</td>
<td>0.134</td>
<td>0.101</td>
<td>0.220</td>
</tr>
<tr>
<td>intercept</td>
<td>1.470</td>
<td>1.460</td>
<td>1.480</td>
</tr>
<tr>
<td>sex</td>
<td>0.030</td>
<td>0.008</td>
<td>0.054</td>
</tr>
<tr>
<td>age</td>
<td>0.002</td>
<td>0.000</td>
<td>0.004</td>
</tr>
<tr>
<td>slope.EU</td>
<td>0.020</td>
<td>-0.130</td>
<td>0.170</td>
</tr>
<tr>
<td>slope.IA</td>
<td>0.010</td>
<td>-0.215</td>
<td>0.201</td>
</tr>
<tr>
<td>precision</td>
<td>80.000</td>
<td>68.200</td>
<td>91.300</td>
</tr>
</tbody>
</table>
Figure 3.3: Distribution of proportional ancestry estimates in the African-American metapopulation using ADMIXMAP. Solid line indicates the theoretical Dirichlet distribution, histogram indicates observed ancestry proportions.
Tests for allelic association of each genotype with each phenotype were performed and results of these tests are shown in table 3.6, along with the results for ANOVA. Shown are Bayesian P values of the score tests. The null hypothesis tested is no effect of the alleles or haplotypes in a regression analysis with individual admixture and covariates as explanatory variables. We observed good concordance in results obtained with the classical methods and the Bayesian methods. The most significant results were observed for BMI and Obesity coded as binary variables with the locus rs3317. HDLc shows an association with rs16383 using the Bayesian method but not with the classical method. The locus rs1042602 shows a suggestive association with SBP. In addition, the Bayesian method detects suggestive association between SBP and locus rs3317. Tests for association of ancestry with the phenotypes show a strong association between the locus rs3317 and BMI for all three ancestry axes. Positive association is observed with proportional African ancestry and negative association is observed with proportional European and Indigenous American ancestry (P<0.0005 in each case). Locus rs1042602 shows a suggestive positive association with African ancestry for DBP and SBP (P = 0.005) and a negative association with European ancestry (P = 0.0059 for DBP and P = 0.0048 for SBP). No ancestry associations were detected for other loci and phenotypes. None of the loci showed significant deviation from HWE.

**Discussion**

**Individual ancestry and population structure**

We have estimated ancestral proportions and tested for ancestry-phenotype correlations in two population samples of African Americans. The two samples are from two different locations, one from Bogalusa, Louisiana and the second from Philadelphia, Pennsylvania. Several phenotypes including SBP, DBP, BMI, blood HDLc and
Triglycerides were measured in these subjects. Proportional ancestry estimated using 34 AIMs showed no significant difference between estimates from the two locations. Based on similar levels of proportional ancestry and the fact that similar ancestral populations are likely to have contributed to both populations, it was possible to combine the samples from two locations for phenotypic analysis.

In the combined sample we have detected admixture stratification as evidenced by correlation in ancestry estimates obtained with different marker sets, and excess of allelic associations between unlinked markers. Ancestry estimates in this sample appear to be in general agreement with those that have been reported previously for other African American populations (Parra et al. 1998; Parra et al. 2001; Shriver et al. 2003; Reiner et al. 2005). It is interesting that in this sample proportional ancestry does not differ between the two geographic locations which was an important aspect that allowed for combining of samples. However it should also be pointed out that we only looked at estimates of individual ancestry and not group estimates. Previous studies have shown that proportional European ancestry is ~ 20% in different locations from US, with somewhat higher European ancestry in Northern and Western US (Chakraborty et al. 1992; Parra et al. 1998; Destro-Bisol et al. 1999; Hoggart et al. 2003). Mean estimates of individual ancestry in the sample are very similar with the two methods used for inferring ancestry. While the maximum likelihood methods depend acutely on the exact ancestral populations and alleles frequencies used, the Bayesian methods have been proposed to take into account the uncertainly associated with such assumptions and to deal with incompleteness of the data arising from missing genotypic information. In general we have used sufficiently large ancestral population samples for estimating allele frequencies and included multiple source populations for each ancestral group. Some concern may yet remain as to the variation among African source populations having contributed to variation within African American samples. Studies have specifically shown that the diversity within African populations (Tishkoff and Williams 2002; Kittles and Weiss 2003) will be important to consider when studying complex phenotypes. However, it has also been shown that markers with high allele frequency difference between Europeans and West Africans, such as ones that have been used in this study, show limited variation.
within Africa (Collins-Schramm et al. 2002). We have used 545 African individuals from Nigeria, Sierra Leone, Central African Republic, Ghana, Congo, Zambia and Uganda for estimating ancestral allele frequencies for this analysis. Nevertheless, more efforts should be made to identify other ancestral populations and to use an appropriate composition of ancestral individuals. It has been suggested that the Bayesian MCMC methods are more appropriate for analyses such as these (McKeigue 1998; McKeigue 2000; Pritchard et al. 2000; Patterson et al. 2004), since these methods do not use fixed values of allele frequencies for estimating ancestry. Bayesian methods are able to iteratively use information from all members of the admixed population and ancestral populations to make inferences for any one individual. In addition, it is possible to use non-informative priors in many cases to overcome problems arising from missing data.

As seen in Shriver et al., 2003 the mean estimates obtained with both methods show remarkable concordance. This could mean that for this sample, the marker panel and admixture model being used has the maximum information. Additional markers will need to be typed for extracting additional information from this sample. The plots of ancestry distribution from the Bayesian analysis show that the data fits the model well, as seen from the comparison of expected Dirichlet distribution to the observed distribution of ancestry in the sample, in Figure 3.3. In absence of such diagnostics, it is difficult to test the reliability of the model assumptions. The same cannot be said for the ML estimates since no definite diagnostics were available to check the model assumptions.

**Ancestry component-phenotype associations:**

BMI levels show significant association with proportional IA ancestry in this sample of African Americans. An association between BMI and African admixture has previously been shown in another sample of African Americans (Fernandez et al. 2003), but no association between BMI and ancestry was shown in a small sample of Puerto Rican women from New York City (Bonilla et al. 2004a). We did not observe an association between SES and BMI, although it is possible that the measure of SES that
we have used is not adequate for detecting any meaningful effects. This is a limiting aspect of the study. Using years of education as a crude measure of SES provides only limited information. Unless better estimates of social, environmental and behavioral factors that affect obesity are included in the analysis, possible confounding by these factors will not be accounted for sufficiently. Nevertheless our results provide additional evidence for the role of genetic determinants underlying obesity in this combined African-American sample. In most analyses in African Americans, investigators have focused on proportional African ancestry in the sample (Fernandez et al. 2003; Gower et al. 2003; Reiner et al. 2005), with the unstated assumption that risk alleles for the phenotypes under consideration are inherited from African ancestors. This is a valid assumption as seen in our results for locus specific effects, where the locus rs3317 shows strong positive association with African ancestry and negative association with European ancestry. But, an overall ancestry-phenotype association may not be the same as single locus specific effects and indeed high risk alleles can come from one or all three ancestral populations. BMI as a measure of obesity is a quantitative trait encapsulating many complex biological functions, each of which are probably influenced by many different genes. Some of these genes are likely to be the high risk alleles inherited from one ancestral population and some from another ancestral population. In addition, functional alleles from one ancestral population may interact differently with other genes in the admixed sample. We have tried to estimate an overall effect of ancestry on the phenotype assuming that such overall effects will reflect a general tendency, which may or may not be true. Further investigation is required to fully understand this apparent incongruity in our results. We found no evidence of overall ancestry associations with HDLc and triglycerides in this sample although we did observe locus specific effects in HDLc levels. This could be due to the fact that overall metabolism and levels of these serum lipids have a more universal causal effect rather than one conditioned by population differences. On the other hand, it could also be that there are more subtle unmeasured effects that affect these phenotypes which we have not been able to detect in this study.

We observe significant negative associations between proportional EU ancestry and BP measures. Hypertension has been used as a classic example of a phenotype where
high-risk alleles in African Americans assumed to be inherited from African ancestors. This assumption is the basis of a recently reported admixture mapping study of hypertension, where loci showing higher proportional African ancestry were identified as ones signaling possible causative alleles (Zhu et al. 2005). While this is a possible explanation, not observing a positive West African-hypertension association does not conclusively prove that no risk alleles have been inherited from the high-risk population. It may be that alleles inherited from the low-risk ancestral population confers some amount of protection or resistance to susceptibility and this could explain the significant negative association between European ancestry and BP. It should be noted that BP in itself is a complex quantitative phenotype affected by many different genes (see for example OMIM # 145500 which lists 15 different chromosomal regions that show associations with essential hypertension) and complex gene-environment-behavioral interactions. We have attempted to investigate a specific interaction that is likely to be regulated and modified by numerous other factors that have not been measured optimally in this study. We have seen suggestive evidence of ancestry on BP and BMI, which partly confirm a genetic basis for these phenotypes. Additional studies are required to replicate these results. A better designed study, using a larger AIM panel and with more precise and well defined measures of environmental and SES variables in conjunction with phenotypic measures, needs to be set up if ancestry estimates are to be fully utilized for decomposing the sources of variation that underlie a complex phenotype.

Locus specific effects on phenotypes

Several loci show associations with the phenotypes studied in this project. After applying Bonferroni correction to multiple tests, the significant P value for each test is lowered to 0.001 for all comparisons. Unlike in some previous studies in which there were numerous associations between genotypes and skin pigmentation, which were subsequently found to be the results of confounding due to population stratification (Shriver et al. 2003; Bonilla et al. 2004a), we have found fewer such associations and
most of these associations persisted even after controlling for variation in ancestral proportions. We also detected one association only after adjustment for ancestry indicating that not controlling for ancestry may also lead to false negatives. It was also interesting to observe the concordance between results obtained with the classical and Bayesian methods. In the Bayesian method a complex model is fit to the data, which includes estimates of ancestry as one of the predictors, but the inference procedure is different from the classical method. A specific hypothesis of no association is tested at each locus and a score and Bayesian P value is calculated for each locus (Hoggart et al. 2003). In general, loci identified to be significant only by the Bayesian method are shown to have suggestive effect using classical methods (for example see locus rs16383 and HDLc in table 3.6). Two loci were found to show significant associations that were not identified by the classical methods: rs16383 for HDLc and rs1042602 for DBP.

Identification of loci using admixture mapping implies that that either the loci are causative agents or more appropriately that the loci are in admixture linkage disequilibrium (ALD) with true causative agents. Our selection of markers was based on ancestry informativeness, rather than candidate genes for complex traits. Some of the AIMs are in genes that are candidate genes in human skin pigmentation (rs6058017, rs1042602, rs2228478, rs1800404). We provide a discussion on some of our AIMs for which there is evidence for linkage by admixture mapping. The most significant association we observed is between BMI and the locus rs3317. This is a SNP in C5orf18: (Locus NP_005660), mutations in which have been associated with familial adenomatosis polyposus and colorectal cancer. This gene is ~ 1.4 Mbp from the gene STARD4 (START domain containing 4, sterol regulated) which is involved in cholesterol biosynthesis and uptake (Soccio et al. 2002; Soccio et al. 2005). Locus rs16383 which shows an association with HDLc is an insertion/deletion polymorphism in the gene encoding a thyroid autoantigen (Chan et al. 1989).

The locus rs1042602 is a SNP in the tyrosinase gene which is a gene previously shown to be responsible for normal variation in skin pigmentation and has measurable effects in skin pigmentation differences between European and African populations (Shriver et al. 2003). We observed suggestive evidence of association of this locus with
Interestingly a previous study has reported an association between this locus and SBP in an African American cohort (Reiner et al. 2005). In an early study on hypertension it was shown that intravenous injections of tyrosinase lowers blood pressure in hypertensive rats and dogs and inactivates rennin, angiotonin and adrenalin (Schroeder and Adams 1941). In addition it has been shown in clinical and epidemiological studies that vitamin-D regulation affects the rennin-angiotensin system that affects blood pressure (Reviewed by Li 2003) and vitamin-D synthesis is associated with skin pigmentation (Murray 1936; Loomis 1967; Holick et al. 1987). It is possible that as yet unidentified gene-gene interactions could result in such a strong association between this locus and blood pressure. Further research is required to confirm this association and investigate any underlying biological basis.

Two loci in the same general region on chromosome 1 show an association with blood pressure. Locus rs6003 shows an association with SBP and locus rs16626 shows an association with hypertension as a binary variable. The locus rs6003 is a SNP in the gene $FXIII\ B$, which encodes coagulation factor XIII B subunit, the last zymogen to become activated in the blood coagulation cascade. It was recently shown that risk of non-fatal myocardial infarction is modified by interactions between allelic variants within $FXIII\ B$ and estrogen use (Reiner et al. 2003). A gene for Pseudohypoaldosteronism type II has been mapped to the chromosomal location 1q31-32 (Mansfield et al. 1997) which is in the same chromosomal location as loci $FXIII\ B$ and rs2065160. It is highly possible that there are putative causative loci in close ALD with ones that have been identified; if not that these are themselves causative loci and further research is needed to understand the exact mechanism by which these loci influence complex traits. Locus rs16626 is on the p arm of chromosome 1 and only hypertension coded as a binary variable shows an association with this phenotype. A previous study has identified several mutations in the gene $ECE1$ located in the region 1p36.1 and proposed this as a candidate locus for blood pressure regulation in humans (Funke-Kaiser et al. 2003).
Conclusion

This study represents a preliminary investigation of complex phenotypes using admixture. Further research is required to replicate and validate these findings in other samples. To fully utilize the potential of the admixture mapping method we would require to type in tens of thousands of markers (McKeigue 2005), which it is likely to be possible soon, given that many such markers have been identified (Shriver and Parra 2000; Smith et al. 2001; Shriver et al. 2003; Shriver et al. 2004; Smith et al. 2004; McKeigue 2005; Shriver et al. 2005) and genotyping costs have lowered. In addition, and perhaps more importantly, we need better characterization and estimates of environmental and behavioral factors that affect complex outcomes. Additional research needs to focus on gene-environmental interactions since it is likely that these reactions play an important role in determining the final outcome. Complex outcomes are “complex” because several factors are involved and typically a study like this one focuses on one particular aspect, in this case genetics. Ancestry-phenotype associations are expected to help decompose the sources of variation that contribute to the phenotype, but even that may not be as straightforward. While studies on skin pigmentation and ancestry associations (Shriver et al. 2003; Bonilla et al. 2004) are encouraging, we do not expect many other phenotypes will show correlations with ancestry that are as strong as the skin pigmentation correlation. Nonetheless, when studying a population with known history of admixture, controlling for variation in genetic heterogeneity is necessary. Ancestry estimates provide one measure of heterogeneity that can be incorporated into an analysis fairly easily. To utilize admixture the study design requires specific populations and markers for effectively modeling the underlying variation in ancestry. By identifying an association between proportional ancestry and BMI and proportional ancestry and BP, this study demonstrates that ancestry is an important determinant that explains a proportion of the variation underlying such complex phenotypes. Several phenotypes have been tested in this project and more than one hundred fifty tests have been performed to look for locus specific associations. Thus multiple testing issues are important considerations in this case. Bonferroni correction has been applied in general
although there are long standing arguments about the merits of "correcting" p-values for multiple testing among Statisticians (Brown and Russell 1997). In the ADMIXMAP program a Bayesian P value is reported and users need to decide how the tests can be interpreted. A more commonsense approach, supported by rigorous arguments, is a Bayesian one. The evidence (likelihood ratio) in favor of a true association at given effect size is estimated as

Prior odds X (statistical power) / p-value

Thus even a p-value of $10^{-4}$ would not be very convincing, if the prior odds of an association at this locus were $< 10^{-3}$, and the statistical power of the study to detect this association was $< 10\%$. Based on this argument, the loci identified in this analysis are suggestive overall and require additional tests in different samples for further confirmation. Overall interpretation of such associations requires care and caution and oversimplification of results is a fallacy that investigators need to be aware of.
References


Flegal KM (1997) Is an inverted weight-height index a better index of body fatness? Obes Res 5,93S(abstr)
causal or confounding association because of population stratification? Hum Genet 110:553-560


Chapter 4

Measurement of individual ancestry, description of admixture structure and using individual ancestry estimates to study group differences among different US populations from the Family Blood Pressure Program

Introduction

The late 1400s that marked the beginning of the European Colonial period, brought together in the new world, populations from several different continents. Thus, many populations (like African Americans, Cubans, Mexican Americans and Puerto Ricans) within the US have genetic roots on more than one continent. Studying genetic admixture is important because unless controlled for, variation in individual ancestry estimates can confound analyses of genetic association. On the other hand, the genetic structure of admixed populations can be used to study and localize genes underlying population differences in disease risk (Chakraborty and Weiss 1986; Chakraborty and Weiss 1988). While early efforts used group admixture estimates (Long et al. 1991; Williams et al. 1992), recent studies have estimated individual ancestry proportions (Williams et al. 2000; Fernandez et al. 2003; Gower et al. 2003; Halder and Shriver 2003; Shriver et al. 2003; Bonilla et al. 2004a; Bonilla et al. 2004b; Shriver et al. 2005) and used these as covariates in genetic association studies to control for the confounding caused by admixture. Estimates of individual ancestry further provide important insights into the stratification inherent in such populations as well as the admixture dynamics of the population (Pfaff et al. 2001). Stratification can arise in a population through several mechanisms including drift, assortative mating and admixture. Stratification as generally defined indicates presence of subgroups within a population. In populations with a history of admixture, stratification will arise as a consequence of the admixture process resulting in variation in ancestry proportions among individuals and if unaccounted for
will lead to confounding. Thus careful examination of admixture structure in admixed populations is warranted.

A recent report (Tang et al. 2005a) examined the genetic structure of five populations from the US and from Taiwan, who were part of the Family Blood Pressure Program (FBPP). The authors used genetic clustering methods to show that self identified race/ethnicity can be used to classify individuals into distinct non overlapping groups and there is little confounding when cases and controls are sampled from the same self identified racial/ethnic and geographic groups. They further demonstrated that in the samples that they have used such confounding would not have affected the studies of hypertension. This follows another report that debated for considering such discrete classification to be useful in medical research (Risch et al. 2002). Race or ethnicity is clearly an important aspect to consider when studying many different phenotypes, since there are established and documented differences in phenotypic expression and severity of the disease between different populations (Weiss et al. 1984; Stern 1995; Nguyen 2003; Nesbitt 2004; Nesbitt et al. 2004; Uchino et al. 2004; Bolton and Wilson 2005; Freedland and Isaacs 2005; Hughes et al. 2005; Polite and Olopade 2005). However, discrete racial categorization provides only a very limited summary of the genetic structure in populations with known history of admixture. Variation in individual ancestry can and should be used when possible to get a better description of variation between populations. In addition, when studying complex traits like type II diabetes, obesity and hypertension, which are characterized by complex gene-gene; gene-environment interactions, associations between proportional ancestry and phenotypes can be used to explicitly test the hypothesis of a genetic basis for population differences of the trait (Weiss et al. 1984; Chakraborty and Weiss 1986; McKeigue et al. 2000; Fernandez et al. 2003; Shriver et al. 2003).

The FBPP was established as a multicenter network to investigate the genetics of hypertension. While the primary focus of FBPP has been identification of genes that cause hypertension using sib-pair linkage analysis, the genome spanning panels of microsatellite loci that have been genotyped in these samples provide a valuable resource for various other analyses. The presence of undetected population stratification has been
implicated as one factor that leads to an excess of false positives in genetic association studies. In populations of mixed ancestry, such as African Americans and Mexican Americans, population stratification arises through the process of admixture itself and can be understood as interindividual variation in admixture proportions. One way to control for such stratification is to include individual admixture estimates in the analytical model. The long range linkage disequilibrium created when previously separated populations admix provides a powerful resource for performing genome-wide mapping, requiring fewer markers and has recently been applied for detecting genes that may cause hypertension (Zhu et al. 2005).

We have estimated individual ancestry in populations from three networks within FBPP and have investigated variation in ancestry levels. We tested for admixture stratification within populations and observed associations between proportional ancestry and some disease related phenotypes in these populations. These are some of the same base populations studied by Tang et al., (2005a). We have focused only on three populations, i.e. European Americans, African Americans and Mexican Americans. Estimates of individual ancestry provide important information related to the genetic structure of the populations that goes beyond discrete categorization whether using the concept of “Races or through genetic clustering (Wilson et al. 2001; Rosenberg et al. 2002; Bamshad et al. 2003; Tang et al. 2005a). Indeed, we have observed that individuals self identifying as being from different populations show overlapping distributions when proportional ancestry is measured. Several studies have shown that major genetic variation in human populations is captured by descriptions of ancient (rather than recent) geographic origin of populations (Rosenberg et al. 2002; Tang et al. 2005b). This does not mean that such categorical descriptions are adequate for understanding within or even between group variations, especially those that may be important for understanding disease risk. In fact, as our analyses show, such categorization may provide an inaccurate picture of populations. Racial/Ethnic group affiliation is a complex cultural process, where social norms and attitudes play a significant role. Studies have also shown that genetic variation within populations are not necessarily discrete, depending on sampling strategies (Serre and Paabo 2004; Shriver et al. 2005).
Using two different approaches we have estimated individual ancestry in 3178 unrelated individuals who self identify as European-American, African-American or Mexican-American. We have used a maximum likelihood method (Hanis et al. 1986) for estimating individual ancestry using previously published multilocus genotypes (Rosenberg et al. 2002) for estimating ancestral allele frequencies. As a complimentary approach we have used a Bayesian method as implemented in the computer program STRUCTURE 2.1 (Pritchard et al. 2000; Falush et al. 2003) to estimate individual ancestry in the same populations. We have also performed computer simulations to explore the reliability of individual ancestry estimates using these two methods for marker panels like the one analyzed in this study, for a similar sample size.

Materials and Methods

Population samples

We used multilocus genotypes for individuals from three populations that have been ascertained as part of the FBPP for admixture analysis. The FBPP is a multicenter network focused on the genetics of hypertension and related phenotypes that was established by the National Heart, Lung and Blood Institute (Williams et al. 2000; 2002; Hunt et al. 2002; Wu et al. 2002). It includes four component networks: GenNet, GENOA, HyperGEN and SAPPHiRe. GenNet, GENOA and HyperGEN ascertained from European-American and African-American families in addition GENOA also ascertained from Mexican-American families from Starr County, Texas. GenNet sampled African-American nuclear families from Maywood, IL and European-American families from Tecumseh, MI, through the identification of a young – middle aged proband with elevated blood pressure. GENOA sampled African-American sibships from Jackson, MI, where both siblings have essential hypertension. European-American families in GENOA are from Rochester, MN, and consist of siblings with essential hypertension and other family members who are not hypertensive. The Mexican-American samples in GENOA
are from Starr County, TX and were selected based on affected sibships. Unaffected family members were used as controls. The samples in HyperGEN were also selected based on hypertensive status in siblings but used unrelated individuals as controls. The European-American persons in HyperGEN are from Salt Lake City, UT, Minneapolis, MN and Framingham, MA. African-American persons are from Birmingham, AL and Forsyth County, NC. The SAPPHIRe network recruited Japanese American, Chinese and Chinese American sib-pairs, one group concordant for hypertension and one group discordant for hypertension, from the San Francisco Bay area, Hawaii and Taiwan (2002; Province et al. 2003; Rao et al. 2003). Individuals were genotyped for >400 autosomal microsatellite loci at the Marshfield Center for Genetic Research and have since been used for several studies (Province et al. 2000; Schork et al. 2000; Hunt et al. 2002; Wu et al. 2002; Province et al. 2003; Thiel et al. 2003; Barkley et al. 2004; Zhu et al. 2005) on hypertension and related phenotypes. For the present analysis, we selected a random sub-sample of unrelated individuals from three networks, GenNet, GENOA and HyperGEN, without considering their disease/phenotypic states, since our primary aim was to understand the genetic variation within and among population samples. This was achieved by choosing one individual from each family for whom we had genotypic and phenotypic data. Our total sample size consists of 1277 European Americans, 1519 African Americans and 413 Mexican Americans.

Ancestral allele frequencies and genotypes were estimated using samples of European, West African and Indigenous American populations that are part of the CEPH Human Diversity Panel (Pritchard et al. 2000; Cann et al. 2002; Rosenberg et al. 2002) and include 109 French and Italians, hence forth referred to as EU to designate putative European ancestors; 69 individuals from various parts of Western Africa (Yorubans from Nigeria, several Bantu and Mandenka from Senegal) hence forth referred to as WAF to designate putative West African ancestors, and 50 Pima and Maya individuals from Mexico, hence forth referred to as IA to designate putative Indigenous American ancestors.
**Genotyping**

Genotypes and allele frequencies for ancestral populations were obtained from the CEPH Human Diversity Panel available at the Center for Medical Genetics at Marshfield (http://research.marshfieldclinic.org/genetics/). These individuals have been typed for a standard panel of 511 microsatellite markers, spaced on average at 10 cM (Screening set 10). All the network samples were genotyped for between 372 and 411 microsatellite markers (Screening sets 8, 9). In this study we used the panel of 311 microsatellite markers that overlap between all network samples and ancestral population samples.

Marker information content for ancestry was estimated as composite $\delta_c$ values which is an extension of biallelic $\delta$, the difference in allele frequencies in two populations, to a multiallelic situation. Following (Shriver et al. 1997) $\delta_c$ is defined as:

$$\delta_c = 1/2 \times \sum_{i=1}^{n} |f_{iA} - f_{iB}|$$

where $f_{iA}$ and $f_{iB}$ are the frequencies of the $i$th allele in the two populations, A and B, being compared at a locus.

**Sample characterization and data analysis**

Allele frequencies and exact tests for HWE in the samples from each network were performed using Genepop software (Rousset and Raymond 1995). All standard statistical tests were done in Minitab version 13 or SPSS version 10 and Windows Excel.

**Estimation of Individual Admixture:**

We stipulated a trihybrid model of admixture between the three ancestral populations EU, WAF and IA, for all populations and individual ancestry was estimated using two methods. We first used a maximum likelihood (ML) method (Hanis et al.
1986) to estimate proportional ancestry using allele frequencies from three ancestral populations. Briefly, the maximum likelihood method computes the highest probability of observing a marker genotype, given ancestral allele frequencies at that locus and admixture proportions from each ancestral population. The admixture proportion that maximizes the probability of obtaining the measured genotype is the ML estimate (MLE) for the individual. Summing over log (MLE) across all loci combines multilocus information to generate an overall estimate of admixture for an individual. Perl scripts were written to calculate individual MLE, (MLIAE) and are provided in Appendix E.

The second method we used is a hierarchical Bayesian MCMC model implemented in the Program STRUCTURE 2.1 (henceforth referred to only as STRUCTURE) (Pritchard et al. 2000; Falush et al. 2003). To estimate individual ancestry we used the “Admixture model” which allows for partitioning genotypes of an individual into separate populations to reflect diverse origins from multiple populations as required. In addition we included all three ancestral populations (as recommended by the author, to augment classification of individuals), and used the POPFLAG option which identifies the individuals from the ancestral populations as representative ancestral subjects. We stipulated separate alpha parameters (parameter for the Dirichlet distribution that characterizes distribution of admixture proportions in the sample). By specifying separate alpha for each population we allowed for variation in the underlying distribution of admixture proportions in each sample. We assumed that allele frequencies are uncorrelated between populations. Within each network populations were grouped by different self identified racial/ethnic affiliation and location. We analyzed individuals from each self described race/ethnic group from each network in separate analyses and also combined individuals from all networks based on similar self identified racial affiliation. Each analysis in STRUCTURE was performed with an initial burn in of 40,000 repetitions followed by an additional 80,000 repetitions for estimating parameters.

To visualize the ancestry estimates we used triangle plots where each vertex of the triangle represents 100% contribution from one ancestral population. Thus, theoretically, individuals that plot at a vertex represent unadmixed individuals with genetic contribution from one ancestral group. Individuals that plot immediately adjacent
to one of the triangle sides are di-hybrid and do not have ancestry from the ancestral population represented by the vertex opposite to that side. For example, individuals plotting along the European and West African axis have negligible Indigenous-American ancestry. Although the triangles plots are set up to allow the representation of three ancestral components in two dimensional spaces, individuals with the same admixture proportions plot on top of each other and thus a single point on the plot may represent multiple individuals. We have thus used histograms to further elaborate how ancestry is distributed in the sample and these histograms can be viewed in conjunction with the triangle plots to understand the admixture structure within populations more completely.

**Tests for admixture structure:**

We have used the Individual Ancestry Correlation test (IACT) (Shriver et al. 2005) to detect presence of admixture stratification in the samples. In this test the total marker panel is divided into two non syntenic sets containing markers on the odd chromosomes and even chromosomes respectively. Individual ancestry is estimated separately with each set of markers and correlation between ancestry estimates obtained with the different subsets is tested. Simulation (Pfaff et al. 2001) and empirical (Shriver et al. 2005) studies have previously shown that this test detects admixture stratification in population samples.

The program STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) was used separately to infer stratification within each population sample. For this test the main aim was to infer presence of stratification within the samples. Each population sample was run separately under an “Admixture” model, without including any ancestral populations and keeping all model parameters the same as described previously. When K $\geq 2$ was more likely than K = 1, we concluded evidence of stratification (Pritchard et al. 2000; Shriver et al. 2003; Bonilla et al. 2004a; Bonilla et al. 2004b).
Ancestry-phenotype correlations:

Several phenotypes were available for the each of the network samples which have been collected using standardized protocols (2002). However, the data set made available to the present investigators, did not have all phenotype information in all population samples. We have tested ancestry-phenotype associations in African-Americans and separately Mexican-Americans using BMI as one phenotype that was available for all individuals studied. In addition we used fasting blood glucose and fasting blood insulin levels that were available for all samples except in the African-Americans from GenNET centers. Estimates of ancestry were used in linear regression models separately for each phenotype. Appropriate covariates were used where applicable. Years of education and smoking status (smokes now/ever smoked vs. does not/never smoked) were used as proxies of SES and behavioral factors, respectively. Age and sex for each individual was used in all models. Based on information related to medication use by participants, individuals’ who are either identified as diabetics or taking antidiabetic medication were excluded from phenotype analysis. All models were separately run using ML ancestry estimates and STRUCTURE ancestry estimates. Since African-American women have greater tendency to be overweight and be hypertensive (Hill et al. 2001; Hall et al. 2003; Tilghman 2003; Nesbitt and Victor 2004), all ancestry-phenotype associations in African-Americans were independently repeated using only female subjects and separately using only male subjects. BMI was transformed to inverse BMI (Flegal 1997) and glucose and insulin were log transformed to normalize the data and henceforth all references to these phenotypes indicate the transformed variables.

Simulations

To further explore the differences in ancestry proportions estimated using ML and STRUCTURE, we performed computer simulations of the admixture process. Simulated data sets of multilocus genotypes were generated using allele frequency estimates from the ancestral populations, i.e., the EU, WAF and IA from the CEPH Human Diversity
Panel. The general algorithm used is as follows: starting with a panel of allele frequencies from multiple population samples, individuals $X_1$, $X_2$, $X_3$…$X_i$ are generated from specified ancestral populations 1, 2, 3….i “n” generations in the past. The individuals generated in the first step of the program all have 100% ancestry from one of the ancestral populations. The numbers of individuals in the first generation who are from one ancestral population correspond to the expected genealogical ancestry proportion in the final simulated population. Thus,

$$X = 2^n$$

Where $X$ is the total number of individuals generated in the first step of the program and $n$ refers to the number of generations before present that the simulation started. For example, to simulate an individual with genealogical ancestry proportions at 0.75 EU and 0.25 WAF, the simplest model would be to start by simulating three EU grandparents and one WAF grandparent. Thus, four individuals are generated in the first step of the program; three have 1.0 EU genealogical ancestry and one 1.0 WAF genealogical ancestry. In the second step of the program, mating between individuals in the first generation is simulated, by iteratively choosing two individuals, without replacement, from this generation, and using alleles at each locus in these “parents” for generating locus specific genotypes of one “offspring”. At each locus chromosomal segregation is simulated by choosing one allele from each parent. Thus, following Mendel’s law each parent has 50% chance of passing any particular allele to an offspring. Two individuals are generated in the second generation. One of these individuals has both parents who have EU ancestry, and this individual is of 1.0 EU ancestry. The second individual in the second generation has one EU parent and one WAF parent and is thus expected to have 0.5 ancestry from each population. Once individuals in the second generation have all been generated, these serve as parents for the next generation. In the final step, mating is simulated between the two individuals in the second generation following the same scheme for choosing alleles from each parent, as in the first generation. One allele at each locus is chosen from each “parent” and combined to generate the individual with expected 0.25 WAF and 0.75 EU ancestry. Additional details
for the simulation scheme are provided in Chapter 5 and the code is provided in Appendix C of this thesis.

It is possible to control the number of generations of the simulation by changing the initial numbers of individuals, as defined in equation 1. So, for example, to start a simulation that goes back 5 generations, we specify 32 ancestors and to start a simulation going back 10 generations we specify 1024 individuals. Our algorithm also allows us to vary the number of generations for the same expected ancestral proportions. For instance, in our first example above, we have simulated one individual with 0.75 EU and 0.25 WAF ancestry by starting with four grandparents, two generations in the past. Alternately, a second model can be specified by starting with 24 EU ancestors and 8 WAF ancestors five generations in the past to generate one individual with 0.75 expected EU ancestry and 0.25 expected WAF ancestry. We have used genealogical ancestry or the proportion of ancestors of an individual that are from one ancestral population as a starting point to generate our simulated samples. This measure corresponds to one expectation of the proportion of genome that can be inherited from one ancestral population. However, due to segregation of alleles at each generation, the actual or realized proportions of alleles inherited by an individual may be different from this expectation. Since the computer programs are designed to estimate an individual’s segregation ancestry (or the proportion of alleles inherited due to the segregation process) it is important to obtain an estimate of the exact proportion of alleles from each population. Especially when the markers being used are not unique to a population or when multi allelic markers are used, assigning ancestry to an allele copy is difficult. Since this is precisely what each of the programs is attempting to do, we devised a method that enabled us to track alleles as they are passed through successive generations and obtain an estimate for the segregation ancestry of each person. This is done by tagging alleles from each ancestral population in the first generation, i.e. first step of the simulation, to indicate population of origin for that allele. Since most of the markers used in this study are not fixed in any ancestral populations, deciphering the ancestral origin of an allele is otherwise not possible. By tagging the alleles we were able to track the alleles from each ancestral population and obtain specific estimates of ancestry (defined by
proportion of alleles in an individual that is inherited from each ancestral population) in
the final individual. The difference between genealogical ancestry and the final
proportion of alleles that an individual inherits is accounted for by segregation of alleles
in each generation. Our algorithm can generate multiple unrelated individuals who have
the same ancestral proportions, in a single iteration. While each individual in this set has
the same genealogical ancestry, each of them has been affected differently by segregation
and we use this estimation of segregation ancestry against which to compare computed
estimates of genomic ancestry made using STRUCTURE and with ML. To generate
samples of individuals with different ancestral proportions, different numbers of ancestral
individuals have been specified and the program was run separately for each model. The
general algorithm is implemented in a perl script, sim_sample_tag.pl that was written for
this purpose and is attached in Appendix C.

We used the general simulation algorithm to generate individuals 200 individual
each belonging to the three ancestral populations with no admixture (i.e. 1.0. ancestry
from a population). We next generated samples of individuals of mixed ancestry. The
tagged allele counts were used for estimating ancestry in each individual and served as an
estimate of the segregation ancestry as described above. The multilocus genotypes
generated by the program were used separately in ML and STRUCTURE under the same
models as used previously for real samples. Significant departures between expected and
observed values of ancestry estimates were examined using standard statistical tests.

Results:

Marker characteristics:

The marker panel used in this network is a standard panel used for linkage
analysis and was selected to cover the genome uniformly. It should be noted that these
markers were not selected specifically for ancestry information content. Total ancestry
information as measured by pair wise $\delta_c$ is 122.09 between Indigenous Americans and
West Africans, 106.54 between West Africans and Europeans and 90.37 between European and Indigenous Americans. Corresponding mean $\delta_c$ values are 0.39, 0.34 and 0.29 per each marker. Figure 4.1 shows the distribution of marker $\delta_c$ values for each pairwise comparison. As seen from this figure, $\delta_c$ values for most markers are between 0.2 and 0.7 (84% of WAF-EU, 85% of WAF-IA and 72% of EU-IA markers). This panel has the least ancestry information for distinguishing between EU and IA ancestry, as reflected by the lower total and average $\delta_c$ values, which may have some impact on individual ancestry estimates made using these panels.

---

**Figure 4.1:** Distribution of composite delta between all pairwise populations in the ancestral population. $\delta$ is highest between West Africans (WAF) and Indigenous Americans (IA) and lowest between Europeans (EU) and Indigenous Americans.
Estimates of proportional ancestry:

Table 4.1 shows the mean estimates of individual ancestry for each sample under a three way model of admixture between EU, WAF and IA ancestors, separately using ML and Table 4.2 shows the estimates for STRUCTURE. Since there are three different European-American and three different African-American samples, we calculated estimates for each sample separately, as well as the total sample of all African Americans and all European Americans. In STRUCTURE, separate simulations were used for estimating ancestry in European-American, Mexican-American and African-American samples. In addition to the combined European-American and African-American samples, we analyzed samples from each population within each network separately.

Table 4.1: Mean ± Standard Deviation of proportional ancestry estimates in FBPP samples using Maximum Likelihood

<table>
<thead>
<tr>
<th>Population (N)</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA (1247)</td>
<td>0.07 ± 0.05</td>
<td>0.90 ± 0.05</td>
<td>0.03 ± 0.03</td>
</tr>
<tr>
<td>EA1 (214)</td>
<td>0.06 ± 0.04</td>
<td>0.91 ± 0.05</td>
<td>0.03 ± 0.03</td>
</tr>
<tr>
<td>EA2 (425)</td>
<td>0.07 ± 0.05</td>
<td>0.91 ± 0.05</td>
<td>0.02 ± 0.02</td>
</tr>
<tr>
<td>EA3 (608)</td>
<td>0.08 ± 0.05</td>
<td>0.89 ± 0.05</td>
<td>0.02 ± 0.02</td>
</tr>
<tr>
<td>AA (1518)</td>
<td>0.05 ± 0.04</td>
<td>0.24 ± 0.12</td>
<td>0.71 ± 0.10</td>
</tr>
<tr>
<td>AA1 (278)</td>
<td>0.05 ± 0.04</td>
<td>0.23 ± 0.10</td>
<td>0.71 ± 0.10</td>
</tr>
<tr>
<td>AA2 (542)</td>
<td>0.05 ± 0.05</td>
<td>0.24 ± 0.12</td>
<td>0.71 ± 0.12</td>
</tr>
<tr>
<td>AA3 (698)</td>
<td>0.05 ± 0.04</td>
<td>0.23 ± 0.09</td>
<td>0.72 ± 0.09</td>
</tr>
<tr>
<td>MA (413)</td>
<td>0.36 ± 0.12</td>
<td>0.59 ± 0.13</td>
<td>0.05 ± 0.02</td>
</tr>
</tbody>
</table>

IA: Proportional Indigenous American ancestry; EU: Proportional European ancestry; AF: Proportional West African ancestry
AA: African Americans in combined sample; AA1: African Americans from GenNet; AA2: African Americans from GENOA; AA3: African Americans from HyperGEN
EA: European Americans in combined sample; EA1: European Americans from GenNet; EA2: European Americans from GENOA; EA3: European Americans from HyperGEN
MA: Mexican Americans from GENOA
Table 4.2: Mean ± Standard Deviation of proportional ancestry estimates in FBPP samples using STRUCTURE

<table>
<thead>
<tr>
<th>Population (N)</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA (1247)</td>
<td>0.06 ± 0.03 #* 1</td>
<td>0.91 ± 0.05 #*</td>
<td>0.04 ± 0.02 #* 2</td>
</tr>
<tr>
<td>EA1 (214)</td>
<td>0.004 ± 0.01 *</td>
<td>0.99 ± 0.012 *</td>
<td>0.003± 0.01*</td>
</tr>
<tr>
<td>EA2 (425)</td>
<td>0.06 ± 0.031</td>
<td>0.90 ± 0.029 *</td>
<td>0.04± 0.02*2</td>
</tr>
<tr>
<td>EA3 (608)</td>
<td>0.07± 0.031</td>
<td>0.89 ± 0.031 *</td>
<td>0.043± 0.02*2</td>
</tr>
<tr>
<td>AA (1518)</td>
<td>0.04 ± 0.01 #* 1</td>
<td>0.17 ± 0.1 #*</td>
<td>0.82 ± 0.9 #*</td>
</tr>
<tr>
<td>AA1 (278)</td>
<td>0.01 ± 0.01 *</td>
<td>0.20 ± 0.08 *</td>
<td>0.79 ± 0.08 *</td>
</tr>
<tr>
<td>AA2 (542)</td>
<td>0.004 ± 0.006 *</td>
<td>0.15 ± 0.09 *</td>
<td>0.85 ± 0.09 *</td>
</tr>
<tr>
<td>AA3 (698)</td>
<td>0.004 ± 0.006 *</td>
<td>0.18 ± 0.08 *</td>
<td>0.82 ± 0.08 *</td>
</tr>
<tr>
<td>MA (413)</td>
<td>0.26 ± 0.16 *</td>
<td>0.74 ± 0.16 *</td>
<td>0.004 ± 0.01*</td>
</tr>
</tbody>
</table>

IA: Proportional Indigenous American ancestry; EU: Proportional European ancestry; WAF: Proportional West African ancestry
AA: African Americans in combined sample; AA1: African Americans from GenNet; AA2: African Americans from GENOA; AA3: African Americans from HyperGEN
EA: European Americans in combined sample; EA1: European Americans from GenNet; EA2: European Americans from GENOA; EA3: European Americans from HyperGEN
MA: Mexican Americans from GENOA

* P Value <0.01 for pair wise t-test comparing ancestry estimates obtained with ML and STRUCTURE and applying Bonferroni correction for multiple testing
# P Value <0.01 for Kruskal-Wallis test comparing proportional ancestry between three networks within each racial/ethnic category and applying Bonferroni correction for multiple testing

In the European-American samples we observed small but statistically significant differences in estimates of ancestry obtained using ML and STRUCTURE. The difference between ML and STRUCTURE estimates was the highest in the GenNet network among all European-American samples. The next set of figures show the distribution of individual genomic ancestry in the different European-American samples from the three different networks. Figure 4.2 shows EA samples from GenNET using ML, Figure 4.3 shows samples from GENOA using ML and Figure 4.4 shows samples from HyperGEN using ML. Figure 4.5 shows the GenNET samples using STRUCTURE, Figure 4.6 shows the samples from GENOA using STRUCTURE and Figure 4.7 shows the samples from HyperGEN using STRUCTURE. Histograms of proportional ancestry in each sample are shown with the triangle plots to better illustrate the variation in genomic ancestry within the samples. In all histograms, IA: Indigenous American ancestry, EU: European ancestry and WAF: West African ancestry.
Figure 4.2: Distribution of individual genomic ancestry estimates in 243 European Americans from GenNET estimated using ML
Figure 4.3: Distribution of individual genomic ancestry estimates in 425 European Americans from GENOA estimated using ML.
Figure 4.4: Distribution of individual genomic ancestry estimates in 607 European Americans from HyperGEN estimated using ML.
Figure 4.5: Distribution of individual genomic ancestry estimates in 243 European Americans from GenNET estimated using STRUCTURE.
Figure 4.6: Distribution of individual genomic ancestry estimates in 425 European Americans from GENOA estimated using STRUCTURE
Figure 4.7: Distribution of individual genomic ancestry estimates in 607 European Americans from HyperGEN estimated using ML
We observed a bimodal distribution within the GenNET network as shown in Figure 4.8 (only ML estimates shown). 28 individuals who were categorized as European Americans, have >30% IA ancestry and cluster separately from the remaining individuals. Including these individuals in the larger sample from GenNET has a small but significant effect on average ancestry levels. Including these 28 individuals lowers mean ML EU ancestry in the sample to 0.88 ± 0.091 compared to 0.091± 0.05 in the sample excluding these individuals (N=214) and increases mean IA ancestry to 0.09 ± 0.04, from 0.06 ± 0.04 in the sample excluding the outliers. Individuals with >30% Indigenous Ancestry were excluded from all further analyses.

![Plot of all 242 individuals within GenNET, including 28 with high levels of Indigenous American ancestry, who were not used for stratification or phenotype analysis.](image_url)

We observed low levels of WAF and IA ancestry in all three European-American samples using ML. Using STRUCTURE, the estimates for the GenNet sample was significantly different from the other European-American samples and showed very low and statistically insignificant non-EU ancestry. Using ML, in all analyses except in
GenNet samples, we saw significantly higher proportional IA ancestry compared to proportional WAF ancestry, as determined by 2 sample t-tests and applying Bonferroni correction for multiple testing (P<0.0001 in each case). We tested for difference in ancestry estimates across networks using the Kruskal-Wallis test which is a non-parametric form of ANOVA. In ML analyses the differences were not found to be significant for any of the ancestry axes. However, in STRUCTURE analyses, significant differences were observed which can be attributed to the sample from GenNet which show very low levels of non-EU ancestry compared to the samples from GENOA and HyperGEN. It is also important to note that in the STRUCTURE analyses we did not observe any individuals from the HyperGEN and GENOA samples to have 100% EU ancestry. While using ML, 17% from GENOA and 13% from HyperGEN show 100% EU ancestry. Further explanation and discussion on this observation is provided later. Overall, estimates with STRUCTURE showed lower variance in ancestry levels in all samples.

Individuals were sampled from multiple field centers in the HyperGEN network and Table 4.3 shows the estimates obtained for each of the four field centers. We did not find any significant difference between different field centers, using either the ML or the STRUCTURE estimates in the Kruskal-Wallis test for each ancestry axis. However, we did observe small but statistically significant difference between estimates obtained with ML and those obtained with STRUCTURE.
Table 4.3: Mean ± Standard Deviation of proportional ancestry estimates in European Americans from different field centers within HyperGEN samples

<table>
<thead>
<tr>
<th>Field Center (N)</th>
<th>ML</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IA</td>
<td>EU</td>
<td>WAF</td>
</tr>
<tr>
<td>7 (69)</td>
<td>0.06 ± 0.04</td>
<td>0.91 ± 0.04</td>
<td>0.03 ± 0.03</td>
</tr>
<tr>
<td>8 (177)</td>
<td>0.08 ± 0.05</td>
<td>0.89 ± 0.05</td>
<td>0.03 ± 0.02</td>
</tr>
<tr>
<td>9 (189)</td>
<td>0.07 ± 0.05</td>
<td>0.90 ± 0.05</td>
<td>0.03 ± 0.02</td>
</tr>
<tr>
<td>10 (182)</td>
<td>0.08 ± 0.05</td>
<td>0.90 ± 0.05</td>
<td>0.02 ± 0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field Center (N)</th>
<th>STRUCTURE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IA</td>
<td>EU</td>
<td>WAF</td>
</tr>
<tr>
<td>7 (69)</td>
<td>0.04 ± 0.02</td>
<td>0.89 ± 0.03 *</td>
<td>0.06 ± 0.02 *</td>
</tr>
<tr>
<td>8 (177)</td>
<td>0.04 ± 0.02 *</td>
<td>0.88 ± 0.03</td>
<td>0.07 ± 0.03 *</td>
</tr>
<tr>
<td>9 (189)</td>
<td>0.05 ± 0.02</td>
<td>0.89 ± 0.03 *</td>
<td>0.06 ± 0.02 *</td>
</tr>
<tr>
<td>10 (182)</td>
<td>0.04 ± 0.02</td>
<td>0.89 ± 0.03 *</td>
<td>0.07 ± 0.02 *</td>
</tr>
</tbody>
</table>

* P Value <0.01 for pair wise t-test comparing ancestry estimates obtained using ML and STRUCTURE and applying Bonferroni correction for multiple testing

The distribution of ancestry estimates in the African-American samples from the three networks using ML and STRUCTURE are shown in the different figures. Figure 4.9 shows distribution of genomic ancestry in the African Americans from the GenNET sample estimated using ML. Figure 4.10 shows the African Americans from the GENOA sample using ML and Figure 4.11 shows the sample from HyperGEN using ML. In all cases individual genomic ancestry shows wide variation with significant European and some Indigenous American ancestry. Figure 4.12 shows the distribution of ancestry estimates in the GenNET sample obtained using STRUCTURE. Figure 4.13 shows the STRUCTURE estimates in the GENOA sample and Figure 4.14 shows the estimates in the HyperGEN sample. Tables 4.1 and 4.2 show the mean estimates obtained in each network and the combined sample using ML and STRUCTURE respectively. The
estimates obtained with the two methods were significantly different for all three ancestry axes in all samples, with ML showing higher proportional EU and IA ancestry and lower proportional WAF ancestry compared to STRUCTURE. The mean proportional ancestry calculated using ML did not differ between networks and was 71% on average which is similar to that reported previously (Parra et al. 1998; Zhu et al. 2005). All samples also have small but significant proportions of IA ancestry (P<0.0001 in each case when tested against a null hypothesis of 0 IA ancestry and correcting for multiple testing). With STRUCTURE, proportional IA ancestry was significantly lower (less than 1% on average) in each sample while using ML the estimates were 0.05 on average for all networks separately and in the combined sample. In addition, Kruskal-Wallis tests on STRUCTURE estimates showed significant differences between networks. When more than two groups are being compared, Kruskal-Wallis test examines equality of means for all samples simultaneously and is rejected if at least one of the means differ from the others. However, it does not specifically test if any one group is different from the others or if all of the groups differ from each other. Hence, additional pairwise t-tests comparing two samples at a time (and applying Bonferroni correction) were done to test which groups were different. These tests showed that while difference between GENOA and HyperGEN samples are lower and insignificant, difference between GenNet sample and those from each of the two other networks was significant.
Figure 4.9: Distribution of ancestry estimates and proportional ancestry in 278 African Americans from GenNET using ML.
Figure 4.10: Distribution of ancestry estimates and proportional ancestry in 542 African Americans from GENOA estimated using ML.
Figure 4.11: Distribution of ancestry estimates and proportional ancestry in 698 African Americans from HyperGEN using ML.
Figure 4.12: Distribution of ancestry estimates and proportional ancestry in 278 African Americans from GenNET using STRUCTURE.
Figure 4.13: Distribution of ancestry estimates and proportional ancestry in 542 African Americans from GENOA using STRUCTURE.
Figure 4.14: Distribution of ancestry estimates and proportional ancestry in 698 African Americans from HyperGEN using STRUCTURE.
Even though the means differed significantly, we observed high correlation in estimates with the two methods for proportional ancestry in all samples, as shown in Table 4.4. Typically correlations were highest for the two major ancestry axes, WAF and EU. As seen from the previous triangle plots, IA ancestry was estimated to be significantly lower with STRUCTURE than using ML. Both methods show the extensive variation in genomic ancestry among individuals within samples from each network.

Table 4.4: Correlation between proportional ancestry estimates obtained with ML and STRUCTURE

<table>
<thead>
<tr>
<th>Sample</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA</td>
<td>0.59</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>EA1</td>
<td>0.4</td>
<td>0.46</td>
<td>0.41</td>
</tr>
<tr>
<td>EA2</td>
<td>0.83</td>
<td>0.79</td>
<td>0.64</td>
</tr>
<tr>
<td>EA3</td>
<td>0.5</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>AA</td>
<td>0.43</td>
<td>0.82</td>
<td>0.85</td>
</tr>
<tr>
<td>AA1</td>
<td>0.646</td>
<td>0.918</td>
<td>0.949</td>
</tr>
<tr>
<td>AA2</td>
<td>0.733</td>
<td>0.926</td>
<td>0.944</td>
</tr>
<tr>
<td>AA3</td>
<td>0.861</td>
<td>0.826</td>
<td>0.662</td>
</tr>
<tr>
<td>MA</td>
<td>0.927</td>
<td>0.916</td>
<td>0.535</td>
</tr>
</tbody>
</table>

Shown are Spearman’s correlation coefficient \( \rho \). All P values < 0.0001 and are significant

AA: African Americans in combined sample; AA1: African Americans from GenNet; AA2: African Americans from GENOA; AA3: African Americans from HyperGEN

EA: European Americans in combined sample; EA1: European Americans from GenNet; EA2: European Americans from GENOA; EA3: European Americans from HyperGEN

MA: Mexican Americans from GENOA

Although multiple field sites within Genoa and HyperGEN recruited African Americans, based on our selection scheme of randomly choosing individuals, we only have samples from two field sites within HyperGEN. Mean estimates of ancestry within each field center are presented in Table 4.5. ML and STRUCTURE estimates both differed significantly for the samples from the two field sites.
Table 4.5: Mean ± Standard Deviation of proportional ancestry estimates in African Americans from different field centers within HyperGEN samples

<table>
<thead>
<tr>
<th>Field Center (N)</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 (226)</td>
<td>0.05 ± 0.04</td>
<td>0.26 ± 0.107 #</td>
<td>0.69 ± 0.11 #</td>
</tr>
<tr>
<td>11 (472)</td>
<td>0.05 ± 0.04</td>
<td>0.22 ± 0.094</td>
<td>0.73 ± 0.01</td>
</tr>
<tr>
<td><strong>STRUCTURE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 (226)</td>
<td>0.005 ± 0.011*</td>
<td>0.21 ± 0.09 #*</td>
<td>0.79 ± 0.09 #*</td>
</tr>
<tr>
<td>11 (472)</td>
<td>0.003 ± 0.002*</td>
<td>0.17 ± 0.08*</td>
<td>0.83 ± 0.07*</td>
</tr>
</tbody>
</table>

*Estimated to be significantly different with ML and STRUCTURE
# Significantly different between field centers

Triangle plots of ancestry estimates and histograms of distribution of genomic ancestry estimates for the Mexican-Americans in GENOA are shown in Figure 4.15 using ML and in Figure 4.16 using STRUCTURE respectively. Table 4.1 and Table 4.2 show mean estimates of ancestry obtained with both methods. We observed high correlations in estimates obtained with the two methods, although there were also significant differences (See Tables 4.1, 4.2 and 4.5). Both methods show extensive variation in ancestry estimates in this sample, with large dispersion along the European-Indigenous American axis. Although proportional WAF ancestry is small, it is still significant (P<0.0001) in both ML and STRUCTURE analyses.
Figure 4.15: Distribution of ancestry estimates and proportional ancestry in 413 Mexican Americans from GENOA using ML.
Figure 4.16: Distribution of ancestry estimates and proportional ancestry in 413 Mexican Americans from GENOA using STRUCTURE.
When we compared the estimates of Mexican Americans to those of the European Americans from each of the networks, we found significant overlap between Mexican-American individuals with high EU ancestry and European-American individuals with high IA ancestry (See Figure 4.17). The figure shows that for some individuals self identified race/ethnicity is a confounder, since genetic ancestry may vary extensively between individuals, causing individuals who are genetically similar to identify with different populations based on cultural and social practices. We compared ancestry estimates between the two sexes and observed no significant difference in proportional ancestry between males and females using either ML or STRUCTURE (Data not shown for any of the samples).

Figure 4.17: Plot of 50 European-American individuals with highest Indigenous American ancestry (blank circles) from the GENOA network and 50 Mexican American individuals with lowest Indigenous American ancestry (dots) from the combined European-American sample.
Tests for stratification

Variation in ancestry estimates within samples does not indicate presence of population stratification in itself since some of the variation could be due to measurement error. We used two methods that specifically test for admixture stratification within populations. First we used the Individual Ancestry Correlation (IAC) test where the entire marker panel was split into two groups, with one group containing all markers on odd chromosomes (186 total) and the other containing markers on the even chromosomes (125 total). Each panel was used separately for estimating individual ancestry in the samples. In the absence of stratification no correlation is expected in the admixture estimates obtained with each marker set separately (Pfaff et al. 2001). Table 4.6 shows the results of this test which was conducted with ML estimates and Table 4.7 shows results with STRUCTURE estimates. Since we tested for three ancestry axes separately we used Bonferroni correction to adjust for multiple testing in each sample. As seen in the table, there is significant correlation in the admixture estimates in all the African-American and the Mexican-American samples, for both ML and STRUCTURE, which confirms presence of admixture stratification in these populations. Applying Bonferroni correction for multiple testing we lowered the significance level for individual tests to 0.01 to achieve an overall significance of 0.05 for each sample.

Table 4.6: Results of Individual Ancestry Correlation Test using Maximum Likelihood

<table>
<thead>
<tr>
<th>Sample</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA</td>
<td>0.08 (0.04)</td>
<td>0.082 (0.05)</td>
<td>0.04 (0.03)</td>
</tr>
<tr>
<td>EA 1</td>
<td>0.06 (0.05)</td>
<td>0.08 (0.24)</td>
<td>0.08 (0.02)</td>
</tr>
<tr>
<td>EA 2</td>
<td>0.02 (0.42)</td>
<td>0.002 (0.94)</td>
<td>0.03 (0.27)</td>
</tr>
<tr>
<td>EA 3</td>
<td>0.04 (0.07)</td>
<td>0.08 (&lt;0.001)</td>
<td>0.006 (0.079)</td>
</tr>
<tr>
<td>AA</td>
<td>0.02 (0.02)</td>
<td>0.45 (&lt;0.001)</td>
<td>0.5 (&lt;0.001)</td>
</tr>
<tr>
<td>AA 1</td>
<td>0.01 (0.78)</td>
<td>0.33 (&lt;0.001)</td>
<td>0.42 (&lt;0.001)</td>
</tr>
<tr>
<td>AA 2</td>
<td>0.02 (0.51)</td>
<td>0.34 (&lt;0.001)</td>
<td>0.38 (&lt;0.001)</td>
</tr>
<tr>
<td>AA 3</td>
<td>0.001 (0.97)</td>
<td>0.42 (&lt;0.001)</td>
<td>0.48 (&lt;0.001)</td>
</tr>
<tr>
<td>MA</td>
<td>0.63 (&lt;0.001)</td>
<td>0.60 (&lt;0.001)</td>
<td>0.17 (&lt;0.001)</td>
</tr>
</tbody>
</table>

Shown are Spearman’s correlation coefficients and associated P values. Significant values in bold
Table 4.7: Results of Individual Ancestry Correlation Test using STRUCTURE

<table>
<thead>
<tr>
<th>Sample</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA</td>
<td>0.12 (0.03)</td>
<td>0.2 (0.08)</td>
<td>0.03 (0.19)</td>
</tr>
<tr>
<td>EA1</td>
<td><strong>0.2 (0.001)</strong></td>
<td><strong>0.13 (&lt;0.001)</strong></td>
<td>0.03 (0.32)</td>
</tr>
<tr>
<td>EA2</td>
<td>0.05 (0.03)</td>
<td>0.02 (0.4)</td>
<td>0.003 (0.89)</td>
</tr>
<tr>
<td>EA3</td>
<td>0.03 (0.3)</td>
<td><strong>0.11 (0.001)</strong></td>
<td><strong>0.12 (&lt;0.001)</strong></td>
</tr>
<tr>
<td>AA</td>
<td>0.15 (0.003)</td>
<td><strong>0.52 (&lt;0.001)</strong></td>
<td><strong>0.62 (&lt;0.001)</strong></td>
</tr>
<tr>
<td>AA1</td>
<td>0.14 (0.03)</td>
<td><strong>0.65 (&lt;0.001)</strong></td>
<td><strong>0.58 (&lt;0.001)</strong></td>
</tr>
<tr>
<td>AA2</td>
<td><strong>0.17 (&lt;0.001)</strong></td>
<td><strong>0.56 (&lt;0.001)</strong></td>
<td><strong>0.56 (&lt;0.001)</strong></td>
</tr>
<tr>
<td>AA3</td>
<td>0.12 (0.002)</td>
<td><strong>0.45 (&lt;0.001)</strong></td>
<td><strong>0.63 (&lt;0.001)</strong></td>
</tr>
<tr>
<td>MA</td>
<td>0.81 (&lt;0.001)</td>
<td><strong>0.81 (&lt;0.001)</strong></td>
<td><strong>0.04 (0.12)</strong></td>
</tr>
</tbody>
</table>

Shown are Spearman’s correlation coefficients and associated P values. Significant P values in bold.

For both tables above EA= all European Americans; AA= all African Americans, MA = Mexican Americans AA1: African Americans from GenNet; AA2: African Americans from GENOA; AA3: African Americans from HyperGEN EA1: European Americans from GenNet; EA2: European Americans from GENOA; EA3: European Americans from HyperGEN, MA: Mexican Americans from GENOA.

Both ML and STRUCTURE analyses show evidence for weak but significant stratification in the European-American sample from HyperGEN ($R^2 = 0.08$, $P<0.0001$ with ML and $R^2=0.11$, $P<0.0001$ with STRUCTURE), although the European-American samples from GENOA and the combined sample does not show evidence of population stratification. Using the estimates from STRUCTURE only, we also found evidence of stratification in the GenNet samples ($R^2 = 0.13$, $P<0.001$). The correlations were much lower in the European-American samples, compared to that seen in the African-American and Mexican-American samples (See Tables 4.6 and 4.7).

In a second test we used STRUCTURE to test for stratification within the samples. For these tests we ran STRUCTURE under the admixture model without including any ancestral populations. The algorithm is STRUCTURE attempts to find evidence for subpopulation in samples by testing for departures from HWE and linkage equilibrium (Pritchard et al. 2000). All other model parameters were the same as used previously. STRUCTURE did not detect stratification in any of the network population
samples, indicating that in samples within networks there is no evidence for population substructure. No evidence of substructure was detected even after the analyses were repeated for the combined European-American and African-American samples.

In conclusion, we found evidence for stratification in the populations with known history of admixture using one test. We did not detect stratification in the combined European-American samples using either test, although the IAC test detected weak stratification in the HyperGEN sample (both for ML and STRUCTURE analyses) and the GenNet sample (STRUCTURE analysis only).

**Ancestry-phenotype correlations**

We tested for the evidence of ancestry-phenotype correlations in all samples, using both ML and STRUCTURE estimates of ancestry and measures of BMI, fasting blood glucose (FBG) and fasting blood insulin (FBI). BMI was transformed into inverse of BMI and FBG and FBI was log transformed for normalization. All references to these phenotypes hereon refer to the transformed variables. Separately we tested for evidence of association between SES (measured as years of education) with each phenotype in each population. Sample characteristics and sex specific differences in phenotypes are shown in Table 4.8. We did not have measures for FBG and FBI in the GenNet network.
Table 4.8: Phenotypic characteristics of FBPP samples

<table>
<thead>
<tr>
<th>Population (N)</th>
<th>Age (in years)</th>
<th>EDU (in years)</th>
<th>BMI</th>
<th>FBG(^1) (mg/dL)</th>
<th>FBI(^1) (mg/dL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA (1518)</td>
<td>52.19 ± 13</td>
<td>13.5 ± 4.4</td>
<td>31.1 ± 7.13</td>
<td>114.56 ± 55.75</td>
<td>12.69 ± 16.4</td>
</tr>
<tr>
<td>AA1 (278)</td>
<td>43.87 ± 13.4</td>
<td>13.03 ± 3.6</td>
<td>30.1 ± 7.5*</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>AA2 (542)</td>
<td>59.8 ± 9.2</td>
<td>12.3 ± 0.4</td>
<td>28.9 ± 4.6</td>
<td>114.8 ± 51.9</td>
<td>14.27 ± 19.43</td>
</tr>
<tr>
<td>AA3 (698)</td>
<td>50.7 ± 12.5</td>
<td>13.9 ± 4.2</td>
<td>31.6 ± 7.4*</td>
<td>114.37 ± 58.7</td>
<td>11.61 ± 12.6*</td>
</tr>
<tr>
<td>EA (1247)</td>
<td>56.32 ± 12.5</td>
<td>15.3 ± 3.85</td>
<td>29.82 ± 6.05</td>
<td>102.36 ± 28.1</td>
<td>8.93 ± 7.26</td>
</tr>
<tr>
<td>EA1 (214)</td>
<td>47 ± 14.67</td>
<td>15.4 ± 3.84</td>
<td>29.3 ± 6</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>EA2 (425)</td>
<td>55.62 ± 11</td>
<td>14.95 ± 0.5</td>
<td>30.5 ± 6.3</td>
<td>100.1 ± 27.7 *</td>
<td>9.1 ± 6.92</td>
</tr>
<tr>
<td>EA3 (608)</td>
<td>55.62 ± 11</td>
<td>15.5 ± 3.9</td>
<td>29.55 ± 5.91</td>
<td>103.97 ±28.3</td>
<td>8.8 ± 28.3</td>
</tr>
<tr>
<td>MA (413)</td>
<td>56.6 ± 0.59</td>
<td>7.4 ± 0.25</td>
<td>30.9 ± 5.9*</td>
<td>151.7 ± 68.9</td>
<td>14.7 ± 14.5</td>
</tr>
</tbody>
</table>

Age: Age in years; EDU: years of education; BMI: Basal metabolic rate; FBG: Fasting blood glucose in mg/dL; FBI: Fasting Insulin in mg/dL
EA: European Americans in combined sample; EA1: European Americans from GenNet; EA2: European Americans from GENOA; EA3: European Americans from HyperGEN
MA: Mexican Americans from GENOA
1. Glucose and Insulin measured only in GENOA and HyperGEN samples
* Significant difference between males and females

In the Mexican-American sample years of education correlates positively with proportional EU ancestry (P = 0.032) and negatively with proportional IA ancestry (P = 0.028). Years of education correlates positively with insulin (P = 0.015) and negatively with glucose (P = 0.006). BMI does not show any association with years of education. We used regression models to test for ancestry-phenotype associations using each ancestry axis separately. Phenotypes were adjusted for age, sex, years of education and smoking status. In addition, both FBG and FBI were adjusted for BMI. We did not observe any associations between ancestry and FBI levels in this sample using either ML or STRUCTURE estimates.

BMI shows significant positive correlation with proportional IA ancestry (R\(^2\) = 0.11, P< 0.0001 with ML estimates and R\(^2\) =0.11, P<0.0001 with STRUCTURE
estimates), and significant negative correlation with proportional EU ancestry ($R^2 = 0.11$, $P = 0.001$ with both ML and STRUCTURE estimates). Since we used inverse of BMI as our variable, these results effectively indicate that BMI has a negative correlation with proportional IA ancestry and positive correlation with proportional EU ancestry in this sample. No correlations were detected with proportional WAF ancestry.

FBG shows significant positive correlation with proportional IA ancestry ($R^2 = 0.03$, $P = 0.006$ with ML estimates and $R^2 = 0.02$, $P = 0.02$ with STRUCTURE estimates), and significant negative correlation with proportional EU ancestry ($R^2 = 0.03$, $P = 0.003$ with ML estimates and $R^2 = 0.02$, $P = 0.012$ with STRUCTURE estimates). Table 4.9 shows the results of these correlations along with the slopes. These correlations between ancestry and phenotype would not be detectable in the absence of stratification and our finding further strengthens the evidence for stratification in the Mexican-American sample.
Table 4.9: Associations between ancestry and phenotypes in FBPP populations

<table>
<thead>
<tr>
<th>Phenotype/Population-Method</th>
<th>Ancestry component</th>
<th>$\beta$</th>
<th>$R^2$</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI/MA-ML</td>
<td>IA</td>
<td>0.009</td>
<td>0.11</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>BMI/MA-STRUCTURE</td>
<td>IA</td>
<td>0.007</td>
<td>0.11</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>BMI/MA-ML</td>
<td>EU</td>
<td>0.008</td>
<td>0.11</td>
<td>0.001</td>
</tr>
<tr>
<td>BMI/MA-STRUCTURE</td>
<td>EU</td>
<td>0.007</td>
<td>0.11</td>
<td>0.001</td>
</tr>
<tr>
<td>FBG/MA-ML</td>
<td>IA</td>
<td>0.21</td>
<td>0.03</td>
<td>0.006</td>
</tr>
<tr>
<td>FBG/MA-STRUCTURE</td>
<td>IA</td>
<td>0.153</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>FBG/MA-ML</td>
<td>EU</td>
<td>-0.211</td>
<td>0.03</td>
<td>0.003</td>
</tr>
<tr>
<td>FBG/MA-STRUCTURE</td>
<td>EU</td>
<td>-0.157</td>
<td>0.02</td>
<td>0.012</td>
</tr>
<tr>
<td>FBG/AA2-ML-Females</td>
<td>EU</td>
<td>0.118</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>BMI-AA-ML</td>
<td>WAF</td>
<td>-0.004</td>
<td>0.03</td>
<td>0.055</td>
</tr>
<tr>
<td>BMI-AA-ML-Females</td>
<td>WAF</td>
<td>-0.003</td>
<td>0.07</td>
<td>0.055</td>
</tr>
<tr>
<td>FBG/AA-ML</td>
<td>IA</td>
<td>-0.194</td>
<td>0.1</td>
<td>0.04</td>
</tr>
</tbody>
</table>

BMI: Inverse of BMI; FBG: Log of fasting blood glucose
MA: Mexican Americans; AA: African Americans from all networks combined; AA2: African Americans from GENOA networks
ML: Estimates obtained with Maximum Likelihood only
STRUCTURE: Estimates obtained with Structure analyses only
Females: Where indicated refers to analyses using females only, otherwise analyses were performed on entire sample using sex as a covariate

We did not observe any correlations between ancestry and any of the phenotypes in the European-American samples from the three networks or the combined sample. To understand stratification in the European Americans we will need to consider several European ancestral populations (like British, Irish, Germans and Italians) along with non-European ancestral groups. It is likely that non-European ancestry has very low effect on the biology of the phenotypes that we have studied. It could also be that the effect is present only in few individuals who have higher levels of non-European ancestry compared to other Europeans. But we were not able to detect because of the large sample of European Americans for most of whom there is little effect overall.
For the African Americans we analyzed samples from each network separately and all combined together (adjusting for sex) and then separately using only female subjects. No correlations were detected between BMI and ancestry in the GenNet sample using either ML or STRUCTURE estimates of ancestry. Within GENOA, no correlations were detected for the total sample. In the analysis on females we found suggestive evidence of correlation between FBG and proportional EU ancestry (MLE) \( R^2 = 0.07, P = 0.05 \), but not with STRUCTURE estimates. No correlations were detected in the HyperGEN samples for any of the ancestry axes, either in the combined sample or in analyses of females. Although, we found difference in ancestry estimates between the two field centers in the HyperGEN network, measures of BMI, FBI and FBG did not differ between the two centers and we did not find any ancestry-phenotype correlations wither within each field center or when sexes were analyzed separately within each field center.

When we combined samples of African Americans from all networks, we found suggestive negative correlation between BMI and ML estimates of proportional WAF ancestry \( R^2 = 0.07, P = 0.051 \). This correlation is also observed in the analysis on females \( R^2 = 0.031, P = 0.055 \), but not in males. We also observed correlations between proportional ML estimates of IA ancestry and FBG \( R^2 = 0.1, P = 0.04 \). No correlations were detected with the STRUCTURE estimates.

**Simulation results:**

Although we observed correlations between individual ancestry estimates obtained with two different methods, there were significant differences. Simulations were used to investigate these differences and evaluate the reliability of the individual ancestry estimates obtained with each method. The simulation program sim_sample_tag.pl was used to generate individuals with known ancestry proportions. Mean ancestry estimates for simulated individuals with no admixture are shown in Table 4.10. In all three cases, ancestry was estimated with very little error using both methods. One sample t-tests performed on each axis showed no significant deviation from expected ancestry.
Typically, ancestry from populations which did not contribute to the simulated sample (for example, total non-EU ancestry in simulated Europeans) was less than 0.01. This estimate for “population bias” represents the total admixture of non-contributing populations to the simulated samples. If we had used perfectly ancestry informative markers, i.e. markers where alternate alleles are fixed in ancestral populations, we would expect no population bias. Note that in most cases population bias is very low (>1%) in this sample using either ML or STRUCTURE. The bias is slightly higher in case of WAF ancestry using ML, (mean of 2.4% and standard deviation of 0.002). “Ancestry bias” is a second measure of informativeness and refers to the total ancestry from one ancestral population summed over all populations that do not have contribution from that population. This measure is computed over all populations being included in one study. For example, in this simulation, ancestry bias for EU ancestry refers to the total EU ancestry in unadmixed Indigenous American and West African simulated persons. Ancestry bias is also low (<1%) with least ancestry bias for WAF ancestry.

Table 4.10: Mean ± Standard Deviation of proportional ancestry estimates in simulated samples with 100% ancestry from one population.

<table>
<thead>
<tr>
<th>ML</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
<th>Population Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>0.998 ± 0.005</td>
<td>0.002 ± 0.005</td>
<td>0 ± 0.0004</td>
<td>0.002 ± 0.005</td>
</tr>
<tr>
<td>EU</td>
<td>0.004 ± 0.014</td>
<td>0.996 ± 0.014</td>
<td>0</td>
<td>0.004 ± 0.014</td>
</tr>
<tr>
<td>WAF</td>
<td>0.007 ± 0.015</td>
<td>0.017 ± 0.028</td>
<td>0.976 ± 0.032</td>
<td>0.024 ± 0.043</td>
</tr>
<tr>
<td>Ancestry Bias</td>
<td>0.011 ± 0.029</td>
<td>0.019 ± 0.033</td>
<td>0 ± 0.0004</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IA</td>
<td>0.998 ± 0.001</td>
<td>0.001 ± 0.001</td>
<td>0.001 ± 0.001</td>
<td>0.002 ± 0.002</td>
</tr>
<tr>
<td>EU</td>
<td>0.001 ± 0.0001</td>
<td>0.998 ± 0.099</td>
<td>0.001 ± 0.001</td>
<td>0.002 ± 0.001</td>
</tr>
<tr>
<td>WAF</td>
<td>0.001 ± 0.001</td>
<td>0.001 ± 0.001</td>
<td>0.998 ± 0.001</td>
<td>0.002 ± 0.002</td>
</tr>
<tr>
<td>Ancestry Bias</td>
<td>0.002 ± 0.0011</td>
<td>0.002 ± 0.002</td>
<td>0.002 ± 0.002</td>
<td></td>
</tr>
</tbody>
</table>

Population Bias: for example, non African ancestry in Africans; Most non African ancestry in Africans; Ancestry Bias: for example, African ancestry in non Africans; Least African ancestry in non Africans

The distribution of genealogical ancestry in the simulated samples of admixed individuals is shown in Table 4.11. As seen in this table we used 17 groups of individuals
whose mean genealogical ancestry is shown here. Weighted mean for the genealogical ancestry in the simulated samples is 0.67 WAF, 0.25 EU and 0.05 Indigenous American. The segregation ancestry estimates for these individuals were collected and used as the expected estimates for all comparisons. The mean genomic ancestry proportion for the combined simulated sample of mixed EU, WAF and IA ancestry is shown in Table 4.12. We did not observe any significant difference in means between observed and expected estimates in the ML analyses. In the STRUCTURE analysis, proportional WAF ancestry was not different, however proportional EU ancestry was significantly higher than expected and proportional IA ancestry was significantly lower than expected.

Table 4.11: Distribution of ancestry proportions in simulated sample

<table>
<thead>
<tr>
<th>N</th>
<th>WAF</th>
<th>EU</th>
<th>IA</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>0.94</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>16</td>
<td>0.87</td>
<td>0.13</td>
<td>0</td>
</tr>
<tr>
<td>54</td>
<td>0.84</td>
<td>0.1</td>
<td>0.06</td>
</tr>
<tr>
<td>16</td>
<td>0.84</td>
<td>0.03</td>
<td>0.13</td>
</tr>
<tr>
<td>29</td>
<td>0.82</td>
<td>0.17</td>
<td>0.01</td>
</tr>
<tr>
<td>47</td>
<td>0.73</td>
<td>0.21</td>
<td>0.06</td>
</tr>
<tr>
<td>104</td>
<td>0.68</td>
<td>0.28</td>
<td>0.04</td>
</tr>
<tr>
<td>25</td>
<td>0.68</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td>51</td>
<td>0.63</td>
<td>0.35</td>
<td>0.02</td>
</tr>
<tr>
<td>43</td>
<td>0.63</td>
<td>0.26</td>
<td>0.11</td>
</tr>
<tr>
<td>46</td>
<td>0.58</td>
<td>0.38</td>
<td>0.04</td>
</tr>
<tr>
<td>24</td>
<td>0.57</td>
<td>0.3</td>
<td>0.13</td>
</tr>
<tr>
<td>29</td>
<td>0.5</td>
<td>0.49</td>
<td>0.01</td>
</tr>
<tr>
<td>30</td>
<td>0.5</td>
<td>0.41</td>
<td>0.09</td>
</tr>
<tr>
<td>12</td>
<td>0.37</td>
<td>0.56</td>
<td>0.07</td>
</tr>
<tr>
<td>7</td>
<td>0.27</td>
<td>0.64</td>
<td>0.09</td>
</tr>
<tr>
<td>AVG</td>
<td>0.67</td>
<td>0.25</td>
<td>0.05</td>
</tr>
</tbody>
</table>

N: Number of individuals in each set of simulated individuals. WAF: Mean proportional West African ancestry in sample of size N, EU: Mean proportional European ancestry in sample, IA: Mean proportional Indigenous American ancestry in sample. AVG: weighted average estimate of ancestry for total sample.
Table 4.12: Mean ± Standard Deviation of proportional ancestry estimates in simulated sample of mixed ancestry

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th></th>
<th>STRUCTURE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>0.06 ± 0.002</td>
<td>EU</td>
<td>0.25 ± 0.01</td>
<td>IA</td>
</tr>
<tr>
<td></td>
<td>0.69 ± 0.01</td>
<td>WAF</td>
<td>0.27 ± 0.01*</td>
<td>EU</td>
</tr>
</tbody>
</table>

* Significant difference between estimated ancestry and expected ancestry after adjusting for multiple testing using Bonferroni correction

Correlations between observed and expected estimates of ancestry were high for the WAF axis (for both ML and STRUCTURE) but were lower for EU and IA axes as shown in Table 4.13. In addition, correlation between estimated ML ancestry and expected ancestry was higher for both EU and IA axes than between estimates of STRUCTURE with expected ancestry proportions. Using Kolmogorov-Smirnov tests we compared the distribution of proportional ancestries we did not find any significant departures from expected in the ML analyses as shown in Figure 4.18. In the STRUCTURE analyses, proportional WAF ancestry was not different from that expected, however, proportional EU ancestry is significantly higher (P = 0.0001) and proportional IA ancestry is significantly lower (P = 0.0002) than expected (after adjusting for multiple testing), as shown in Figure 9. Further research is required to determine the biological significance of these departures. For our study this analysis demonstrates that ML estimates are probably more reliable than the STRUCTURE estimates when using the current marker panel.

Table 4.13: Correlation between expected and observed estimates of ancestry

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th></th>
<th>STRUCTURE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>0.64</td>
<td>EU</td>
<td>0.88</td>
<td>IA</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>WAF</td>
<td>0.71</td>
<td>EU</td>
</tr>
</tbody>
</table>

Shown are Spearman’s correlation coefficients. P values were <0.0001 and significant in all cases.
Figure 4.18: KS test comparison percentile plots comparing expected ancestry (solid line) against estimated ancestry (broken lines) using ML (left side panels) and STRUCTURE (right side panels). A) Plots comparing proportional West African ancestry, B) Plots comparing European ancestry, C) Plots comparing Indigenous American ancestry.
Discussion:

Variation in individual ancestry in populations

This study demonstrates the variation in individual ancestry in three populations that are part of the Family Blood Pressure Program and how such variation can be used for understanding group differences. Using large population samples we have shown that the variation in individual ancestry confounds discrete categorization of individuals based on self identified racial/ethnic groups. Our results contradict those from a previous study (Tang et al. 2005a) where clustering methods on the same data showed that self identified race/ethnicity shows good concordance with major genetic clusters. However, in their use of the algorithm implemented in the program STRUCTURE, these authors used a model which does not allow for individuals to have ancestry from multiple populations. When used in this way, the program fits a model where each individual is assumed to originate from a single population (Pritchard et al., 2000). The authors justified that their use of the “no admixture” model was to ensure that populations could be treated as separate groups who may have shared genetic ancestry in the past but have sufficiently diverged since then to be identified as distinct, non overlapping groups (Tang et al., 2005a). They also stated that they did not use individual estimates of ancestry which would only be appropriate for understanding within population variation.

We have investigated genetic variation within populations using estimates of individual ancestry and observed significant overlap in individual ancestry among Mexican Americans and European Americans (see Figure 4.17 for example). While African Americans in this study do not overlap with either of the two other populations, the heterogeneity within this population is quite substantial. Some African American individuals in this sample have more EU and/or IA ancestry compared to other African Americans from the same sample, which is likely to be true for most African American populations. The populations included in this analysis, i.e. European Americans, African Americans and Mexican Americans, all have recent shared history. The European colonial period brought together populations that had previously been separated. This led
to the creation of admixed populations, like the African Americans and Mexican Americans, each with substantial EU ancestry. In addition, intermixing between African and Indigenous Americans have resulted in more complex admixture scenarios (Palmer 1976; Palmer 1991). Sociological studies have investigated the extent of non-EU ancestry in European Americans to show presence of IA and WAF ancestry in populations that identify as “White” (Stuckert 1958; Unrau 1989; Logan and Ousley 2001). A previous study using 34 ancestry informative markers estimated 3.2% IA ancestry and 0.7% WAF ancestry in a sample of European Americans from Pennsylvania (Shriver et al. 2003). Whether or not this admixture has any effect on the disease phenotype is yet to be determined, these evidences argue that further exploration of individual ancestry is needed in self identified European Americans. Of course, greater understanding of variation within European Americans will require better characterization of the variation within Europe and how such variation may have affected phenotypes. Sociological evidence suggests that many light skinned African Americans often “passed” for “White” or even “switched” races and within one or two generations identified as “White” (Mills 1998). Thus, untangling the genetic components of ancestry within and between populations requires careful considerations and using self identification may not provide sufficient information. We have used a three-way admixture model with EU, WAF and IA ancestors. It is important to recognize the shared history that has shaped the genetic structure of these populations. Thus in order to understand the genetic structure of these populations it is important to consider the effects of admixture and use an appropriate model. For part of our analysis we have used the same program used by Tang et al., (2005a), except that we have allowed for admixture in individuals to reflect the population histories as closely as possible and followed the recommendations of the program author, who suggests using the admixture model for accurate estimation of parameters instead of using an artificially restrictive model that does not reflect the histories of these populations and including ancestral populations.

Using two different methods we estimated individual ancestry and observed substantial variation in individual ancestry within all population samples. In addition to the combined African-American and European-American samples, we have investigated
variation in individual ancestry within each network for understanding general trends within populations. Individual ancestry varies extensively among African Americans and Mexican Americans, populations that have recent and probably low levels of continuing admixture. We also observed substantial variation in ancestry estimates in the European Americans, especially with the estimates obtained from the program STRUCTURE. Some of the variation in European Americans may be measurement error, but it is also evident that low levels of non-European ancestry are observed in different populations. Variation in individual ancestry results in overlap in proportional ancestry between individuals who self identify as being from different race/ethnicities (See figure 4.17 for example). We have used estimates of individual ancestry to describe the genetic structure in populations and our results demonstrate that self identified race/ethnicity as a discrete categorical variable is both imprecise and incomplete as a means to describing the genetic structure in populations since there is substantial variation within the racial categories and noteworthy overlap between individuals who self identify as being from different race/ethnicities. It is important to recognize this variation within and between populations in order to understand the genetic basis underlying phenotypes which may show differences between populations.

Markers used for ancestry estimation

We have used a panel of 311 microsatellite markers for estimating individual ancestry Although, these markers were not chosen for ancestry information content, we found that total ancestry information as measured by pair wise $\delta_c$ to be 122.09 between IA and AF, 106.54 between AF and EU and 90.37 between EU and IA. Corresponding mean $\delta_c$ values are 0.39, 0.34 and 0.29, which are comparable to ancestry informative marker panels that have been used for estimating individual ancestry previously (Shriver et al. 2003; Bonilla et al. 2004a; Bonilla et al. 2004b). In this panel we found the least difference in allele frequencies between Europeans and Indigenous Americans, which probably reflects their recent shared common ancestry relative to West African
populations and may partly explain the excess IA ancestry we have observed in the European-American samples in this study. Very few markers in this panel have $\delta_c > 0.5$, and of those markers that do show high $\delta_c$, are ones that primarily distinguish between African and non-African populations (see Figure 4.1). The total number of markers used in this study, 311 microsatellites, is greater than many of the earlier studies on individual ancestry estimation (Parra et al. 1998; Cerda-Flores et al. 2002; Bamshad et al. 2003; Shriver et al. 2003; Bonilla et al. 2004a; Bonilla et al. 2004b; Barnholtz-Sloan et al. 2005) based on which, the estimates of ancestry are expected to be more precise (Chakraborty 1986). But it is important to remember that ancestry information content of markers will affect the overall estimates of ancestry for both of the methods that have been used. To specifically distinguish between the effects of markers from the effects of populations, we simulated individuals who have ancestry from only one population and estimated ancestry using both methods. The results in Table 4.10 show that both methods estimate ancestry reliably, with lower variance in proportional ancestry when using STRUCTURE than using ML with this panel of markers. It would be interesting to estimate proportional ancestry in these samples using well defined and preferably a balanced panel of markers that will differentiate between all populations equally.

Another and probably the most important aspect to remember when using these markers is that these are multi allelic microsatellites. Perfectly informative AIMs would provide the most reliable estimates of ancestry with any method. With SNPs, allele frequency differences intuitively provide more information, since there are only two alleles to consider. Thus, although allele frequency difference provides a good indication of informativeness and has been used for identifying multi allelic AIMs (Smith et al. 2001; Collins-Schramm et al. 2002), information for each allele is difficult to establish. The situation is more confounded for rare alleles which may be present in only the admixed sample but not in the ancestral populations and vice versa. In such situations, ML treats the allele as missing data. Theoretically, Bayesian methods are developed to make use of such data, so that all available information is adequately used. However, the details of how the algorithm in STRUCTURE incorporates this information are not clear and require further analyses.
Methods

We have used two methods for estimating individual ancestry and we found high correlation in estimates of African Americans and Mexican Americans, and moderate correlation in European Americans using the two methods. This indicates that the general trends in ancestry estimates within populations is measured adequately (alternately, they could both be wrong), however, the differences between ancestry estimates with the two methods highlights some issues which need to be considered when inferring individual ancestry. Previous studies on individual ancestry which have used both ML and STRUCTURE observed high correlations in estimates obtained with the two methods (Shriver et al. 2003; Bonilla et al. 2004a; Bonilla et al. 2004b; Reiner et al. 2005; Tang et al. 2005b) to conclude reliability of the estimates obtained. It is important to remember that these estimates were obtained with panels of selected AIMs. One report further showed that ML estimates obtained under a three-way admixture model shows strong correlations with STRUCTURE estimates obtained under a two-way admixture model in a sample of African Americans (Reiner et al., 2005). High correlations obtained with different methods may indicate that either both methods perform equally well for estimating ancestry or that both are systematically erroneous. To investigate the reliability of the estimates we have obtained we performed additional simulations. We have attempted to follow a unified pattern for estimating ancestry, by using the same admixture model (i.e. three way admixture model) and same ancestral populations for both methods. First we used a maximum likelihood method (Hanis et al. 1986), where individual genotypes are used together with ancestral allele frequencies for inferring estimates of individual ancestry. Next we used a Bayesian Markov Chain Monte Carlo Method implemented in the program STRUCTURE (Pritchard et al. 2000; Falush et al. 2003), where genotypes of individuals being investigated are provided along with genotypes of representative ancestral individuals and the program uses these to jointly infer allele frequencies in each subpopulation at each locus and the posterior estimates of individual ancestry following a Bayesian probability model. The main advantage of these Bayesian methods is that they are not restricted by fixed parameters and have the
flexibility of making inferences based on all available data as well as account for missing data. Detailed descriptions of the programs are beyond the scope of this chapter, and are presented in Chapter 5 of this thesis. In the first set of simulations we generated unadmixed individuals and estimated ancestry in these samples. As shown in Table 4.12, mean estimates of ancestry were very close to those expected in the simulated sample, using both methods. However, the variances were lower with STRUCTURE than with ML. We estimated population bias in samples which is a measure of proportional ancestry from a non-contributing population in the sample (for instance total non-African ancestry in simulated 100% Africans) and observed that when using ML, population bias is highest in Africans, indicating that using this marker panel one would expect to find ~2% non-African ancestry in African individuals by chance. Using STRUCTURE, population bias is the same for all populations and is typically <1%. This estimate can also be interpreted as a measure of the marker panel, particularly in conjunction with ML estimates in which proportional ancestry of an individual is estimated irrespective of other individuals included in the study sample. Using STRUCTURE we found population bias to be similar for all ancestry axes, which also reflects the fact that since STRUCTURE uses all individuals to make inferences for any one individual, when a large number of genetically similar individuals are included in the analyses more reliable estimates are obtained. A second measure is Ancestry Bias which reflects total ancestry from a non-contributing population combined over other populations, for example total IA ancestry estimated in EU and WAF samples. Using ML we found that Ancestry bias was the lowest for African ancestry, i.e. one would expect to see least African ancestry in non-Africans. With STRUCTURE, these estimates are again all very similar. These results may be interpreted as a positive attribute for STRUCTURE, since it appears that in unadmixed individuals estimates are closer to those expected. However since our intention is often to estimate proportional ancestry from a sample with varying ancestry proportions, this issue is not adequately explained by this analysis. What we do infer is that overall, in this panel of markers there is a small chance of identifying ancestry from a population when there is no contribution from that population in the samples being studied.
We next simulated multi locus genotypes for individuals of mixed EU, WAF and IA ancestry. Ancestry proportions in our simulated sample are similar to those observed in real African-American populations, with primarily WAF ancestry and substantial EU ancestry. IA ancestry is low in the sample overall, but few individuals have very high IA ancestry. We analyzed the simulated sample the same way as the real samples and compared the estimated ancestry proportions to known ancestry proportions. Estimated proportional AF ancestry deviates very little from known ancestry, using both methods. However, proportional EU ancestry appears to be over estimated and proportional IA ancestry is under estimated using STRUCTURE (see Tables 4.12 and 4.13 and Figure 4.18) in contrast the distribution of ML estimates shows remarkable concordance with the expected distribution of ancestry for all axes. This could be due to the fact that in samples of mixed ancestry, the ancestry that contributes least to the sample is under represented and consequently underestimated by STRUCTURE. For instance, IA ancestry is typically low in African Americans. In our analyses, we have used a sample of only 50 Indigenous Americans as representative of ancestral individuals. But the samples of African Americans are typically much larger (for example, the smallest African American sample from GenNet has a sample size of 278) and these individuals are primarily of EU-WAF ancestry. For those individuals who do have very high IA ancestry, which in our simulated sample was 23% with > 0.1 IA ancestry, this results in overall low proportion of representative IA alleles in the total data that the algorithm starts with. Consequently proportional ancestry from IA is underestimated. In addition, the markers used were not pre selected to be ancestry informative and are least informative for EU-IA distinctions. This confounds the analyses even further. Together, these factors likely contribute to discrepancies observed. Typically, simulations done thus far have focused on a two-way admixture model (Falush et al. 2003; Tang et al. 2005) and such a situation has not been encountered previously. Additional simulations are required to better understand the performance of the STRUCTURE algorithm under different admixture scenarios.

One known source of error with the Bayesian method relates to the inclusion of ancestral individuals along with admixed individuals and typically the program performs more efficiently when representatives of ancestral populations are included (J Pritchard,
2005, Personal Communication). We included the same ancestral individuals (from the CEPH Human Diversity Panel (Rosenberg et al. 2002)) who have been used for estimating ancestral allele frequencies for the ML method. Second the sample sizes of each subpopulation included in the study has an important effect on the final estimate using Bayesian methods as the program uses all individuals, ancestral and admixed samples, to make inferences about one individual. This is where the relative and absolute sizes of the subpopulations become critical [See also ((Tang et al. 2005)]. In our studies, sample sizes for the ancestral populations are much smaller compared to the admixed samples. We have used 109 Europeans, 69 Africans and 50 Indigenous Americans, compared to which our smallest sample size within one network was 214 European Americans from GenNet. When more individuals who are genetically similar are included in the data, the distribution of alleles for that population is much better defined compared to a subpopulation of smaller size. This is apparent from our simulations of unadmixed individuals. In each case we increased the sample size of one population ~ 6X compared to the two other ancestral populations who do not contribute to the sample. For all three simulated ancestral samples, nearly perfect estimates were obtained in each case. This is even more interesting when we compare European and Indigenous American samples because 57% of the markers have $\delta_c <0.3$ between European and Indigenous Americans. While this can possibly lead to more errors in ancestry estimates when analyzing either Europeans or Indigenous Americans, the unadmixed individuals from either of these samples do not show significant ancestry from the other populations. One possible explanation for this is that by including a large number of individuals from one population, we have virtually assured the correct clustering of these individuals into the appropriate representative group.

While ML estimates show good concordance with the simulated data, one needs to be cautious when interpreting these results. We have used two European populations, two Indigenous American populations and individuals from 3 West African populations as ancestors. While these are appropriate samples to start with, they do not reflect the entire range of variations that could and probably have contributed to the ancestry of samples studied. In addition, we have used extant descendants of assumed ancestors.
Drift is a significant evolutionary factor we have not considered. It is possible that allele frequencies in modern descendants of ancestral populations are different from those that originally contributed to the admixed population under study, but using ML, we are unable to consider such changes in the available data. Estimates of ancestry using ML depend critically on ancestral allele frequencies provided, thus small sample size and choice of the ancestral population themselves will have important impacts. Previous study of African Americans from this sample used some of the European Americans from the GenNet study as an ancestral population (Zhu et al. 2005). While this is a reasonable approach, our intention was to also investigate variation within European Americans and hence our choice of ancestral populations.

**Stratification in samples**

The most significant effect of variation in individual ancestry is related to the presence of substructure in population samples. Population substructure can be attributed to several factors including admixture, assortative mating, geographic heterogeneity or drift. In a population with known history of admixture, variation in individual ancestry is the most likely cause for stratification and we have used the term admixture stratification to refer to this phenomenon. Detecting stratification is a difficult problem in association studies and methods have been developed to control stratification (Pritchard et al. 2000; Pritchard et al. 2000; Devlin et al. 2001; Satten et al. 2001). However, methods that rely on detecting substructure based on deviations from HWE and linkage equilibrium (Pritchard et al., 2000) may be underpowered for detecting admixture stratification when such stratification has not lead to the creation of subpopulations. Instead, there is a continuum of ancestry proportions in a sample that could cause confounding in association studies. Because of this continuum in ancestry proportions conventional tests for HWE and linkage disequilibrium are insufficient since within each sample there has been sufficient random mating for achieving HWE (Cerda-Flores et al. 1992). Indeed, in
our tests for HWE we observed <0.005 % of markers deviate from HWE in the admixed populations.

Specific tests that are more appropriate for admixed populations are the individual ancestry correlation test (Shriver et al. 2005), the \( D_i/D_0 \) test (Pfaff et al. 2001; Bonilla et al. 2004a; Bonilla et al. 2004b) that examines the correlation between observed and expected levels of LD and the unlinked marker correlation test (Bonilla et al. 2004a; Bonilla et al. 2004b). Simulation and empirical studies have previously shown that the individual ancestry correlation test detects admixture stratification well (Pfaff et al. 2001), since it relies on variation in individual admixture estimates in the sample. If there is some systematic variation in ancestry proportions in the sample than any ancestry informative marker should be able to detect this variation. Thus, correlation in ancestry estimates obtained with unlinked, non-syntenic marker panels can be used as an indicator of stratification. Using this test we detected presence of admixture stratification in all African American and Mexican American samples. Typically, high correlations were seen in the admixed samples using both ML and STRUCTURE estimates of ancestry. In the combined European American sample we did not observe any evidence for stratification, despite the low levels of variation in ancestry estimates. ML and STRUCTURE both show low correlation within HyperGEN network, and STRUCTURE also detects correlation in the GenNet network for the major axis of ancestry. But the functional implication of this is not clear.

Our purpose was to mainly examine inter-continental admixture effects and we have used an admixture model specifically for that purpose. Typically non-EU ancestry is low in European Americans. However, contemporary European American populations have resulted from admixture between different EU populations like the British, French, Irish, Italians and Germans and a more accurate description of ancestry would involve taking some account of these variations within Europe using suitable ancestry informative markers that distinguish between EU populations. Given the lower genetic diversity within EU populations compared to African populations, finding such markers is going to be difficult. Efforts are currently underway to find markers that will better characterize diversity among European samples (Shriver 2005, personal communication)
and once such markers have been identified will be helpful in better understanding the diversity within European American populations.

**Ancestry-Phenotype correlations**

The effects of population stratification become even more relevant in the presence of observed ancestry-phenotype associations. We have seen significant association between ancestry and phenotypes in the Mexican-American samples. In this sample BMI increases with increasing EU ancestry and decreases with increasing IA ancestry, which is contradictory to trends observed in the Gila River Indian community which has a similar admixture history (Williams et al. 2000). Estimates obtained with both ML and STRUCTURE show this association and it is not immediately clear as to why opposite trends are seen in these two populations. Measures of fasting blood glucose show a positive association with IA ancestry and a negative association with EU ancestry, which is consistent with trends reported previously in other Mexican-American populations (Williams et al., 2000; Chakraborty et al. 1986). The association between ancestry and phenotype indicates a genetic basis for the phenotype; however, interpretation of this association is not straightforward and will require a more detailed examination. Despite the evidence for association of IA ancestry with diabetes (Gardner et al. 1984; Chakraborty et al. 1986; Williams et al. 2000), one report suggests that Spanish ancestry may partly be responsible for higher prevalence of diabetes in a sample of Mexican Americans (Lorenzo et al. 2001). This study also found lower overall adiposity in Mexican American subjects compared to Spanish subjects although central adiposity showed an opposite trend. Overall, the association of BMI and fasting blood glucose with ancestry indicates a genetic basis for these phenotypes in Mexican Americans and additional studies are required to better understand this effect.

Within African Americans, sex is an important factor to consider, since females have higher rates for obesity, hypertension and diabetic nephropathy compared to males (Gillum 1996; Kumanyika 1997; Jones 1999; Crook and Patel 2004). We controlled for sex in our analyses and separately evaluated ancestry-phenotype associations in females.
Preliminary analyses have been reported on association between WAF ancestry and BMI in African American women (Fernandez et al. 2003; Gower et al. 2003). We did not detect any association between ancestry and fasting insulin, after controlling for age, sex, SES and smoking status. Both age and smoking are independently associated with BMI in both the combined sample and in sample of females. After controlling for age, SES and smoking, association between BMI and ancestry is only detected when samples from all three networks are combined and only when ML estimates are used. Further analysis within the sexes showed that the association is present in females and not in males. It is unlikely that geographical difference can be used to explain this observation, since differences in ancestry were typically low between the networks, or between sexes within and between each network. The associations observed with FBG and ancestry is also interesting. Since no measures for glucose and insulin were available for the GenNet network, combined analyses for these phenotypes were done only in subjects from the GENOA and HyperGEN networks. In the network analyses we observed an association between glucose and EU ancestry only in females in the GENOA network. No such associations were observed in the combined sample from all networks. Instead, ML estimates of IA ancestry show an association with age and sex adjusted glucose. Separately this association is observed in the females but not in males. Given that IA ancestry is low in the sample overall, the significance of this association is unclear and further analyses need to be done to investigate this association.

We did not observe any ancestry phenotype associations within European Americans, although these samples have low but significant amounts of non-EU ancestry. This is partly due to the fact that we did not detect strong evidence of admixture stratification in these samples. Even where such stratification was detected, it was low and the present data is insufficient for explaining the functional effects of this stratification. But what is interesting is that comparing average estimates of BMI between the three racial/ethnic groups, we did not find significant difference in BMI levels between African Americans and Mexican Americans, although both differed significantly from European Americans. Discrete categorization based on self identified race/ethnicity does not provide any more information than this. However, our analysis of individual
ancestry shows some interesting aspects and provides additional insight into the biological mechanism that underlies these differences. For instance, we have seen that individual ancestry in Mexican Americans considerably overlaps with that of European Americans. Thus, there are likely some individuals who self identify as European-American but have high proportional IA ancestry. We have found that within Mexican Americans, BMI varies with proportional European ancestry and this provides us additional information regarding the genetics of BMI in this population. But no such association is seen in the European Americans. From our analyses we have not seen any overlap between African Americans and other populations, although there is considerable variation of individual ancestry proportions within African Americans. Thus, there are obvious differences among individuals who self identify as African Americans, with some showing higher EU and/or WAF ancestry than others. Discrete racial categorization assumes homogeneity within groups which is insufficient to account for stratification. Only when we combined individuals across all networks we observed an association between proportional ancestry and BMI, both in the combined sample and in the analyses on females.

In conclusion, we have observed significant variation in ancestry proportions in all the populations we have studied, more in ones with known history of admixture. We have observed that variation in proportional ancestry follows a continuum and for the European-American and Mexican-American samples there is some overlap. We have observed association of BMI and glucose with ancestry in the Mexican-American sample, which confirms the presence of stratification in this sample. Modest associations were also observed in African-Americans. In both populations there is strong evidence of admixture stratification and any association study undertaken in these samples would have to be controlled for variation in individual ancestry. We also find that self identified race/ethnicity provides only limited and in some cases inaccurate description of underlying genetic variation in individuals. Given the significant amount of variation in ancestry proportions both of these populations are excellent candidate for admixture mapping. However, the present marker panel leaves more to be desired in terms of ancestry information content. In addition, to understand the biology of complex traits
effectively, we will need to better characterize phenotypes and consider more specific effects of environmental and behavioral determinants.
References:

use in mapping by admixture linkage disequilibrium. Am J Hum Genet 70:737-750
Flegal KM (1997) Is an inverted weight-height index a better index of body fatness? Obes Res 5,93S(abstr)


Chapter 5

Effects of markers and methods: Practical considerations for estimating individual genomic ancestry

Introduction

Individual estimates of ancestry provide important information regarding the distribution of genetic variation in population samples. As shown in Chapters 2, 3 and 4 of this thesis, variation in admixture proportions results in a continuum of ancestry estimates in a sample, and need not necessarily create discrete subpopulations. Therefore, understanding stratification in an admixed sample requires better understanding of the distribution of ancestral proportions in the sample. Different methods have been proposed for estimating individual ancestry. These include a maximum likelihood method (based on Bernstein’s method and developed later by (MacLean and Workman 1973a; MacLean and Workman 1973b). This method was subsequently used in a study investigating gall bladder disease association with individual ancestry in Mexican Americans (Hanis et al. 1986) and has since then been used extensively for estimating individual admixture proportions in different populations (Williams et al. 1992; Shriver et al. 1997; Parra et al. 2001; Cerda-Flores et al. 2002; Bertoni et al. 2003; Brutsaert et al. 2003; Fernandez et al. 2003; Gower et al. 2003; Shriver et al. 2003; Bonilla et al. 2004b; Parra et al. 2004; Bonilla et al. 2004a; Reiner et al. 2005; Salari et al. 2005; Shriver et al. 2005). More recently, several Bayesian Markov Chain Monte Carlo (MCMC) methods have been proposed as alternatives for estimating individual ancestry that are better able to deal with uncertainties associated with data collection and model specification (McKeigue et al. 2000; Pritchard et al. 2000; Dawson and Belkhir 2001; Anderson and Thompson 2002; Falush et al. 2003; Hoggart et al. 2004; Excoffier et al. 2005). It is important to point out that apart from individual ancestry estimation; research has been carried out on methods...
for estimating group admixture proportions where statistical issues have been dealt with extensively. Since the focus of thesis is individual ancestry estimation and investigating how individual ancestry is distributed in different populations, I have not considered any methods that deal with group admixture estimation.

In the preceding chapters I have used two different methods to estimate proportional ancestry in individuals. One is the maximum likelihood method and the other is Bayesian MCMC method implemented in the program STRUCTURE. Although proportional ancestry estimated using these two methods shows good correlation, assumptions inherent in the methods are likely to result in differences in the estimates of genomic ancestry obtained. It is possible that the statistical differences observed have no biological significance, in which case good concordance in estimates obtained with different methods could be an indicator of reliability of these estimates. But if these estimates are to be used for making inferences more closely related to biology, then understanding such differences will help decide which estimator should be used for making inferences. But they could still both be in error or not very precise.

Apart from methodological issues, the estimates of ancestry will also depend on the ancestry informativeness of the marker panel being used. Previous studies on individual admixture proportions have used smaller panels of markers (typically > 40) (Hanis et al. 1986; Cerda-Flores et al. 1991; Parra et al. 1998; Parra et al. 2001; Cerda-Flores et al. 2002; Bertoni et al. 2003; Gower et al. 2003; Bonilla et al. 2004; Bonilla et al. 2004a; Reiner et al. 2005; Salari et al. 2005), chiefly because of the lack of availability of well characterized ancestry informative marker (AIM) panels. However, with recent efforts in identifying markers informative across different populations, larger and better characterized panels are now available (Smith et al. 2001; Collins-Schramm et al. 2002; Frudakis et al. 2003; Shriver et al. 2004; Smith et al. 2004; Shriver et al. 2005). But as seen from the previous chapters, the combined effect of markers that have been used in conjunction with the method used may have more significant impact on estimates of ancestry.

Studies done to date have focused either on specific methods or on application of one or more of the methods to real data. Proponents of Bayesian methods generally argue
that this is a better approach to deal with real data sets where there are a number of uncertainties including model choice (i.e. which ancestral populations should be used), and allele frequency estimates in different populations and missing data. Although considered computationally not feasible up until a few years ago, these methods are becoming more popular.

The ML model that we have used depends critically on estimates of allele frequencies provided, ancestral populations assumed and is not amenable to handling missing data. However, within the confines of these assumptions the method provides consistent and reliable estimates of ancestry proportions and their standard errors are asymptotically unbiased with an increasing number of marker loci (Edward's 1992). The Bayesian MCMC methods on the other hand require some information on the prior distribution of admixture proportions, which depend on the amount of admixture expected in the sample (Pritchard et al. 2000), although it is possible to use a uniform or non informative prior and update the parameter of interest using the observed data. In a Bayesian framework, the information about parameters of interest is summarized in a 'posterior density'. This is a probability distribution that describes what is known about the value of the parameter after looking at the data. The posterior is proportional to the likelihood times the ‘prior’ (the information which describes what is known before looking at the data). By using a uniform (or flat or non-informative) prior, likelihood dominates the final outcome. With large sample sizes and many AIMs posterior means and 95% confidence intervals obtained with Bayesian methods are asymptotically equivalent to maximum likelihood estimates and 95% confidence intervals (McKeigue 2005). In addition, since the Bayesian methods use both ancestral populations and admixed samples for inferring ancestry for any one individual, misspecification of ancestral populations is of less consequence. Whereas the ML method searches for a global optimum combination of all the parameters, the Bayesian approach aims to explore the space that is consistent with the data, i.e. it does not settle towards a single value.

As discussed previously, either method may perform well in some situations. Numerous studies have been done to develop and understand the classical statistical
methods based on maximum likelihood and least squares (see for example, discussion by Chakraborty, 1986) it is obvious that researchers in this area have made a concerted effort to understand the process of admixture and its implications. Bayesian methods for individual admixture estimation are fairly recent and have garnered considerable interest in keeping with the general resurgence of applying Bayesian techniques to complex biological problems (Beaumont and Rannala 2004). Statistically elegant, Bayesian methods are computationally and statistically intensive. In addition, the Bayesian methods that are available for individual admixture estimation simultaneously make inferences on other parameters as well. For instance, STRUCTURE jointly infers allele frequencies in subpopulations, assigns individuals into clusters and infers proportion of each individual’s genome that is derived from different ancestral populations. In some cases, making joint inferences may not be necessary, in which case excess computational time could be a disadvantage. There could also be situations when violating assumptions of ML would have more serious consequences, for instance, using improper ancestral populations or miss-specifying the admixture model. However, both empirical and theoretical studies are as yet lacking that provide an indication for when one method may be more useful or more applicable than others. Since the aim of these methods is to infer the proportion of alleles that have been inherited from an ancestral population, using suitable markers for making inferences will always be important. The present analyses were undertaken to investigate some specific issues. Assuming that the marker panels being used are sufficiently informative I have investigated 1) whether STRUCTURE provides more reliable estimates of ancestry compared to ML, 2) whether including ancestral samples in STRUCTURE has a significant effect, 3) whether specifying the ‘Popflag’ option in STRUCTURE has a significant effect, 4) whether using independent vs. same alpha parameter in STRUCTURE has a significant effect, 5) whether using a balanced marker panel (one which is equally informative for all ancestry axes) provides more reliable estimates compared to an unbalanced panel.

To answer these questions I have used a sample of simulated individuals of known genomic ancestry. For each individual in the simulated sample I have ascertained ancestry of each allele copy and have estimated proportion of alleles from each ancestral
population. I have separately used ML and several options in STRUCTURE and estimated individual ancestry in each individual and compared these against expected ancestry proportions. These are preliminary studies and provide some important insight into how the two programs function under otherwise ideal situations.

A brief description of the algorithms used in the analyses is provided prior to describing the general simulation scheme.

**Maximum Likelihood model:**

Starting with allele frequencies in putative ancestral populations and marker genotypes in the admixed population, the final aim of this method is to identify the proportion of alleles inherited from each ancestral population. The basic assumption of this method is somewhat “smooth gene flow between ancestral and contemporary generations” (Chakraborty 1986), implying a linear combination of allele frequencies and admixture proportions. This model is briefly described as follows:

Let $p_1, p_2, \ldots, p_k$ denote the frequencies of a specific allele in $k$ populations, and each population contributes a fraction of $m_i$ $(i=1 \ldots k)$ in the formation of an admixed population, the expected frequency of that allele in the admixed population (i.e. contemporary generation) $p_h$ is given by

$$p_h = \sum_{i=1}^{k} m_i p_i$$

Both $p_i$ and $p_h$ can be estimated from data. When there are 2 ancestral populations, i.e. $k=2$, a simple derivation gives the MLE and estimated variance of MLE. From the preceding condition, assuming a di-allelic, di-hybrid scenario, where the admixed population is formed by intermixture between two parental populations, proportions of different genotypes observed in an individual are obtained from the following equations:

$$Pr.(AA) = [mp_1 + (1-m)p_2]^2 \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 1$$

$$Pr.(Aa) = 2[mp_1 + (1-m)p_2] [mq_1 + (1-m)q_2] \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 2$$
\[ Pr. \ (aa) = [mq_1 + (1-m)q_2]^2 \quad \cdots \cdots \cdots \cdots 3 \]

Where AA, Aa and aa represent all possible genotypes at a locus; \( p_1 \) and \( q_1 \) are allele frequencies of alleles P and Q in population 1 and \( p_2 \) and \( q_2 \) are allele frequencies of alleles P and Q in population 2. Since both \( P_1 \) and \( P_2 \) are assumed to be known, the only unknown here is \( m \), the proportional contribution from each ancestral population. To combine information across multiple loci, log likelihoods over all loci are summed assuming that each locus is independent. Since only a single parameter \( m \) is to be estimated, the estimation can proceed over the entire range of possible values of \( m \), i.e. assuming no admixture from an ancestral population (\( m = 0 \)) to assuming 100% ancestry from that ancestral population (\( m = 100 \)). The level of \( m \) where the combined likelihood across loci is the highest is the best estimate of a person’s admixture proportion.

As seen from the above discussion, the ML method we have used only searches for the highest likelihood within the confines of the model (i.e. ancestral populations and frequencies are known without error). Inference based on ML depends largely on asymptotic theory, which demands a reasonably large sample size, independence and various regularity conditions. In addition the problem of multiple tests may become severe when the dimension of parameter space is large. The experiment-wide error rate (type I error) could be difficult to control in a large scale classical problem. For example, when the number of alleles measured is large, the simultaneous inference of \( m_i \)’s for all loci is difficult and computationally intensive (but not impossible). Of course, several classical methods are available for searching the parameter space more efficiently by using algorithms such as Newton-Raphsons or Fisher-scoring methods. Further tests are required for evaluating the effects of numerous loci on estimates of \( m \).

Given the rigidity of this method, it is obvious that any inferences made with ML will depend on the admixture model and the markers used for this analysis. Thus, in situations where there is a lot of uncertainly in the data, this method is unlikely to be reliable. Recently, an extension to this ML method has been developed which incorporates uncertainties related to ancestral populations and ancestral allele frequencies (Tang et al. 2005a). Further analyses are required to understand and evaluate the efficiency of this method.
Bayesian MCMC method implemented in STRUCTURE algorithm

Below is a description of the model in STRUCTURE which specifically allows for admixture in the sample. Using similar notations as in (Pritchard et al. 2000), the basic program is described as follows. Let the vector X denote the observed genotypes in all populations (admixed and ancestral populations if included), Z denotes the (unknown) populations of origin of the individuals, and P denotes the (unknown) allele frequencies in the populations, Q denotes the (unknown) admixture proportions for each individual. These vectors have the following structures.

\[ (x_i^{(i)}, x_i^{(2)}) \] genotype of the ith individual at the lth locus where \( i=1,2,\ldots N \) and \( l=1,2,\ldots L; \)
\[ z_i^{(i,a)} \] population of origin of allele copy \( x_i^{(i,a)} \); (This is important and differs from the model with ‘no admixture’ where inference is based on \( x_i^{(i,a)} \))
\[ p_{lj} \] frequency of allele j at locus l in population k where \( k=1,2,\ldots K \) and \( j=1,2,\ldots J_l \), where \( J_l \) is the number of distinct alleles observed at locus l;
\[ q_k^{(i)} \] proportion of ith individual’s genome that originated from population k.

Having observed these, the Bayesian inference is based on the following model:

\[ P(Z, P, Q | X) \]
\[ \propto P(Z) \ P(P) \ P(Q) \ P(X | Z P Q) \]

In our problems, the primary interest is to estimate the vector Q. The probability model for \((X, Z, P, Q)\) is given by the following equations

\[ P(x_{i}^{(i,a)} = j | Z, P, Q) = p_{z_{i}^{(i,a)} j} \] .................................4
\[ P(z_{i}^{(i,a)} = k | P, Q) = q_{k}^{(i)} \] .................................5

P and Q are assumed to have non-informative proper prior distributions, which is the usual Dirichlet distribution. Independence is assumed among individuals. The vector Q is characterized by a Dirichlet distribution with the parameter \( \alpha \). The prior tends to use all individuals to get a sense of what the ancestry proportions are like. There is a possibility that this assumption could be violated if the true distribution is nothing like a Dirichlet distribution, however, it is not possible to check this assumption directly. In a
separate program Hoggart et al. (2003); Hoggart et al. (2004) have implemented a similar algorithm where they provide an additional graphical comparison of expected distribution to observed distribution of admixture proportions in the sample (see Figure 3.3 in chapter 3). The program provides two options, one in which all individuals are assumed to have the same $\alpha$ parameter, second in which individuals are assumed to have different $\alpha$ parameters. This is an important aspect of the program, since incorrectly specifying a model will have important consequences as shown later.

While it is not possible to directly obtain the distribution of $P(Z, P, Q \mid X)$, the MCMC algorithm will simulate the posterior distribution $P(Z, P, Q \mid X)$, using the conditional distributions described previously. The posterior inference will be based on the simulated sample from $P(Z, P, Q \mid X)$. MCMC refers to a simulation technique that is used to obtain approximate samples from a probability distribution. The idea behind MCMC is as follows: To sample a probability distribution a Markov chain is constructed using standard methods like the Metropolis-Hastings algorithm or Gibbs sampling (which is a special case of the Metropolis-Hastings algorithm for sampling a discrete distribution). A Markov chain is a sequence of random values whose probabilities at a time interval depend upon the value of the number at the previous time. Once the Markov chain has been constructed, a sample of draws is obtained from the target distribution by simulating the Markov chain a large number of times and recording those values. The sampling of the chain constitutes Monte Carlo simulation whose function is to compute an expectation or summarize a given density by sampling the density using a suitable method and then using the sampled draws to compute the expectation or summarizing the density. MCMC samples the probability space by moving around the posterior distribution, and the time that it spends in different parts of the space is proportional to the posterior probability. The Bayesian posterior inference is straightforward on the summary statistics of simulated posterior distributions by MCMC. For example, for the most interesting parameter $Q$, the posterior mean is good point estimation for the true mode (but not for the true mean),

$$E(q_i \mid X) \approx \frac{1}{M} \sum_{m=1}^{M} q_{i}^{(m)}$$
where M is the size of simulated posterior sample given by MCMC. This is the key aspect to remember when comparing estimates from STRUCTURE to those obtained with ML.

The important issue related to MCMC is the transition probability or how the chain moves from one part of the probability space to another. The algorithm in STRUCTURE is designed to move to lower probability parts of the space at an appropriate rate so that the stationary distribution of the Markov chain matches the posterior distribution being estimated. In a well mixing chain each iteration of the MCMC will land at some point on the posterior probability surface whose probability is proportional to the relative probability of each point on the surface. Convergence is achieved when the chain moves into the appropriate parameter space that fits the data.

The primary advantage of the above Bayesian method is that it can handle extremely complex parameter spaces in the ancestry estimation problem. The model has a number of parameters which are influenced by the number of individuals (i.e. experimental subjects) and the large number of alleles at all loci. If all the previously mentioned conditional probability assumptions are met, which are quite straightforward and simple, Bayesian approach will be a feasible choice.

The disadvantage of the Bayesian method is that MCMC algorithm is computational expensive. It usually takes hours or even days to approximately achieve the stationary states, i.e. the true posterior distributions and there is no clear cutoff point in MCMC when the stationary states are arrived at. For using MCMC we still need to derive some necessary conditional distributions based on the model assumption. However from the description of the algorithm (Pritchard et al. 2000) all details for derivation are not articulated. It is possible that additional assumptions of independence are implicitly applied. For example, it is assumed that Z (unknown population of origin) and P (unknown allele frequencies) are independent. It is possible that there are other additional independence assumptions without illustration in the admixture model.
Practical considerations and issues

I have used both ML and STRUCTURE methods for inferring genomic ancestry and in general observed overall concordance in estimates obtained with the two methods. However, there were also significant differences in some cases which merit additional scrutiny. Since the aim of either method is to infer the proportion of alleles inherited by an individual of mixed ancestry, it is possible to use simulated data to investigate different aspects of these methods. A second and more practical reason was to evaluate how the estimates obtained vary under different situations related to the available data (e.g. characteristics of markers used, sample size etc.). Although ML has been used extensively in the literature, using STRUCTURE for estimating ancestry is a relatively new approach and merits detailed investigation. While as a method this is an extremely elegant approach, to fully understand the capabilities of this method, detailed investigation is warranted. Not understanding the details of how STRUCTURE makes inferences or using the algorithm incorrectly may lead to erroneous or misleading conclusions. Methodological issues may be even more confounded by the choice of marker panels. Thus, one aspect of the simulations was to see how the choice of marker panels together with choice of methodology affects estimates of ancestry.

As I have considered a three-way admixture model for all analyses in the previous chapters, I have explored the effects of methods and markers under the same model. I have developed and used a simple population model (described in detail in Methods) in which it is possible to establish and track admixture proportions, measured as proportion of alleles from different ancestral populations in an admixed individual. Since Bayesian MCMC methods can make inferences of admixture proportions even in the absence of ancestral populations this was an important aspect to investigate. In addition, I have explored several parameters within the STRUCTURE algorithm which are likely to have important consequences on admixture estimates. These include the effects of ‘α’ parameter that characterizes the Dirichlet distribution of admixture proportions Q and specifying the ‘POPFLAG’ option which is important for using prior population information to assist in clustering. This work represents preliminary analyses in
comparing the effects of methods and markers on estimates of ancestry. Since both
marker and methods will jointly affect the estimates of ancestry it is important to be able
to fix one variable to investigate the effects of the other. For instance, it is obvious that
ML inferences will be impossible in the absence of ancestral population allele
frequencies. It is not obvious whether using a balanced panel of markers will provide
more reliable estimates of ancestry under a three-way admixture model than an
unbalanced panel. I have used two panels of AIMs, one which is relatively balanced for
information content from all three ancestral populations and one which is not balanced.
Practical reasons for doing so relate to the model of admixture being assumed. Several
studies have assumed that African Americans, a widely studied population group that has
primarily European and West African ancestry (Parra et al. 1998; Smith et al. 2001;
Fernandez et al. 2003; Gower et al. 2003; Patterson et al. 2004; Smith et al. 2004; Zhu et
al. 2005; Tang et al. 2005a). Indeed, marker panels developed for admixture mapping in
African-American populations have been selected to distinguish between European and
West African ancestries with very little consideration to the fact that based on the history
of peopling of the Americas, there is a possibility of some Indigenous American ancestry
in these populations, which may be substantial, depending on where the populations were
sampled. Thus, I have used two relatively large marker panels, a panel of 102 markers
which is balanced for ancestry information content for all three ancestral populations
(European, Indigenous American and West African) and the other a panel of 100 markers
which is chosen to distinguish between Europeans and West Africans only. These
represent idealized situations with respect to markers since very few studied to date have
used >50 AIMs for inferring individual ancestry (Shriver et al. 2005). But these are also
ideal for providing insights into the methodological aspects without additional
confounding of having insufficient marker information.
Methods

Simulation scheme:

I used a three-way model of admixture between Europeans, West African and Indigenous Americans and simulated a sample of mixed individuals using the Perl script Sim_Sample_tag.pl. The basic scheme has been described in detail Chapter 4 of this thesis and the code for the program is provided in Appendix C. Briefly, we start by simulating a set of individuals in generation one who has 100% ancestry from different populations. This is achieved by using allele frequencies from known populations and randomly sampling alleles at each locus to generate the genotype at that locus. Individuals generated at the first iteration of the algorithm are used as ‘parents’ for the next generation. This is achieved by selecting the first two individuals in the array of ancestors from the first generation. ‘Mating’ between the individuals is simulated and an offspring is generated who inherits one allele at each locus from each parent. Thus while each parent randomly passes on one allele at a locus, none can pass on more than 50% of the alleles to an offspring. Once all individuals in the second generation have been generated, these serve as the ‘parents’ for the next generation. Mating is simulated again and the entire process can be repeated for the number of generations stipulated by the researcher. It is possible to vary the genealogical ancestry of individuals, by varying the order of individuals from each ancestral population in the first generation.

Multiple individuals with the same genealogical ancestry can be generated in a single iteration of the program, each individual being generated independently of the others. Although within groups’ genealogical ancestry is the same, segregation of alleles at each generation results in different individuals inheriting different proportions of alleles from the ancestral populations. In addition, it is possible to generate a sample of individuals with varying ancestry proportions and one in which the admixture event has occurred in different generations in the past, by simulating different sets of individuals with different proportions of genealogical ancestry.
The most important aspect of this scheme is that the segregation of alleles in each generation can be tracked precisely. Each allele is tagged in the first generation as being from one ancestral population. Consider that individuals generated in the first iteration of the algorithm all have 100% ancestry from one ancestral population. Thus, by tagging the alleles in the first generation, it is possible to follow and track the ancestral state of each allele through any $n$ number of generations. This is particularly useful since the aim of all individual ancestry estimation programs is to be able to determine the proportion of alleles inherited from the ancestral populations. In the absence of perfectly informative AIMs (i.e. markers for which alternate alleles are fixed in alternate populations) ancestral states cannot be determined with absolute certainty. All loci are simulated as unlinked and follow Mendel’s rule of independent assortment.

The first set of simulations consisted of 200 unadmixed individuals from each of the three ancestral populations. These individuals were simulated by starting with available allele frequencies from each ancestral population and simulating mating for three generations. The final set of 200 individuals was used as the putative ancestral populations for simulating individuals of mixed ancestry. Allele frequency estimates used for ML analyses were derived from the panel of 200 individuals and genotypes for all 600 ancestral individuals were included in STRUCTURE runs where ancestral populations were used. The program was then used to simulate a sample of individuals of mixed European, West African and Indigenous American ancestry of varying proportions. 17 sets of individuals were generated with varying genealogical ancestry proportions as shown in Table 5.1. Within each group, varying numbers of individuals were specified. The overall distribution of ancestry estimates was designed to reflect a real African American population (corresponding to the combined African-American population from the FBPP studies described in Chapter 4). Unadmixed ancestral individuals were simulated 9 generations in the past. Different genealogical ancestries were simulated by varying the proportion of ancestral individuals generated in the first step of the algorithm and by varying the order in which the ancestral individuals were specified.
Table 5.1: Distribution of genealogical ancestry in sample of simulated individuals

<table>
<thead>
<tr>
<th>N</th>
<th>WAF</th>
<th>EU</th>
<th>IA</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>0.94</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>16</td>
<td>0.87</td>
<td>0.13</td>
<td>0</td>
</tr>
<tr>
<td>54</td>
<td>0.84</td>
<td>0.1</td>
<td>0.06</td>
</tr>
<tr>
<td>16</td>
<td>0.84</td>
<td>0.03</td>
<td>0.13</td>
</tr>
<tr>
<td>29</td>
<td>0.82</td>
<td>0.17</td>
<td>0.01</td>
</tr>
<tr>
<td>47</td>
<td>0.73</td>
<td>0.21</td>
<td>0.06</td>
</tr>
<tr>
<td>104</td>
<td>0.68</td>
<td>0.28</td>
<td>0.04</td>
</tr>
<tr>
<td>25</td>
<td>0.68</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td>51</td>
<td>0.63</td>
<td>0.35</td>
<td>0.02</td>
</tr>
<tr>
<td>43</td>
<td>0.63</td>
<td>0.26</td>
<td>0.11</td>
</tr>
<tr>
<td>46</td>
<td>0.58</td>
<td>0.38</td>
<td>0.04</td>
</tr>
<tr>
<td>24</td>
<td>0.57</td>
<td>0.3</td>
<td>0.13</td>
</tr>
<tr>
<td>29</td>
<td>0.5</td>
<td>0.49</td>
<td>0.01</td>
</tr>
<tr>
<td>30</td>
<td>0.5</td>
<td>0.41</td>
<td>0.09</td>
</tr>
<tr>
<td>12</td>
<td>0.37</td>
<td>0.56</td>
<td>0.07</td>
</tr>
<tr>
<td>7</td>
<td>0.27</td>
<td>0.64</td>
<td>0.09</td>
</tr>
<tr>
<td>AVG</td>
<td>0.67</td>
<td>0.25</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Marker Panels:

Two different sets of markers were used in the analysis. Panel 1 consists of 100 AIMs chosen from the Affymetrix 10K Whole Genome Amplification set (Kennedy et al. 2003; Shriver et al. 2005) that distinguish between European (EU) and West African populations (WAF). Genotypes and allele frequencies in these populations were estimated from a panel of 20 Spanish from Valencia, Spain and from 20 Mende from Nigeria (Shriver et al. 2005) which have been used as estimates of European and West
African ancestral allele frequencies respectively. In addition, this panel also has allele frequencies from 20 Nahua and 20 Quechua which were combined as estimates of Indigenous American (IA) allele frequencies. It is notable that there are likely important levels of imprecision in these allele frequencies and that were this panel to be developed as a recommended marker panel verification of frequencies would be critical. However, it is not improper that the allele frequencies be used as representative allele frequencies for our current use here.

Panel 2 consists of 102 AIMs chosen from published allele frequencies available from the Affymetrix 10K WGA array (Shriver et al. 2005), the SNP consortium data base and the HapMap database. Markers were chosen to balance ancestry information content, as measured by allele frequency differences between Europeans, West Africans and Indigenous Americans. 34 AIMs were chosen to be informative for each of the axes of ancestry.

Individual ancestry estimation:

The ML method was implemented using a Perl script MLIAE written for this purpose and provided in Appendix E. Separate analyses were performed using markers in the balanced panel and using markers in the unbalanced panel.

Four different options in the program STRUCTURE were used in separate analyses, each using the balanced and unbalanced marker panels. Since STRUCTURE can use uniform priors in the absence of suitable ancestral populations, the first model in STRUCTURE was chosen to test the effects of not including any ancestral population. In these analyses (denoted henceforth as NP: No Parental populations) the data consisted only of the sample of admixed individuals. In the second model, the effects of including ancestral individuals were tested explicitly. Theoretically, the most reliable estimates of ancestry are provided when appropriate representatives of ancestral populations are included in the analyses by specifying the ‘Usepopinfo’ option. In addition, the ‘Popflag’ option was specified in the model which is the option that indicates whether or not to use the prior population information (as indicated by including the ancestral populations) in
the model (henceforth designated as F: Model with PopFlag). The third model was
specified in which ancestral populations are included but not identified i.e. Popflag option
is not used (henceforth referred to as NF: Not using Popflag). This model was specified
to explicitly evaluate the utility of identifying individuals as being ancestral, as opposed
to treating all individuals as independent cases. The fourth option evaluated the effects of
specifying independent alpha (parameter characterizing the Dirichlet distribution for
distribution of ancestry proportions in the sample. Specifying separate alphas allows the
program to model different Q (ancestral proportions) for each ancestral population. The
alternate model was also used where all populations are assumed to have the same alpha
parameter. Each of the options NP, F and NF were used with either Same (S) or
Independent (I) alphas. Thus, for the STRUCTURE analyses there were a total of six sets
of analyses each using the balanced panel and separately the unbalanced panel. To
specify the different models used in STRUCTURE the codes are as follows:
1. NP-I-B: No ancestral populations, independent alpha, balanced marker panel
2. NP-S-B: No ancestral populations, same alpha, balanced marker panel
3. NP-I-UB: No ancestral populations, independent alpha, unbalanced marker panel
4. NP-S-UB: No ancestral populations, same alpha, unbalanced marker panel
5. F-I-B: Ancestral populations included and identified using Popflag, independent
   alpha, balanced marker panel.
6. F-S-B: Ancestral populations included and identified using Popflag, same alpha,
   balanced marker panel.
7. F-I-UB: Ancestral populations included and identified using Popflag, independent
   alpha, unbalanced marker panel.
8. F-S-UB: Ancestral populations included and identified using Popflag, same alpha,
   unbalanced marker panel.
9. NF-I-B: Ancestral populations included but not identified with Popflag,
   independent alpha, balanced marker panel.
10. NF-S-B: Ancestral populations included but not identified with Popflag, same
    alpha, balanced marker panel.
11. NF-I-UB: Ancestral populations included but not identified with Popflag, independent alpha, unbalanced marker panel.
12. NF-S-UB: Ancestral populations included but not identified with Popflag, same alpha, unbalanced marker panel.

All analyses were performed under a three-way admixture between European (EU), West African (WAF) and Indigenous American (IA) ancestral populations.

**Statistical analyses**

Two main aspects have been compared in these analyses, means of ancestry estimates, correlations between observed and expected ancestry, and distribution of ancestry proportions in the sample. Since estimates of segregation ancestry were available for each individual, all estimates obtained were compared with this expectation. Estimates of square error for each individual was calculated as \[((\text{observed WAF}-\text{expected WAF})^2 + (\text{observed EU}-\text{expected EU})^2 + (\text{observed IA}-\text{expected IA})^2)\] and was averaged for the entire population to obtain an estimate of mean square error for the model. Means for each ancestry axis were estimated separately and were tested against the expected mean using 2 sample t-test and applying Bonferroni correction for multiple testing. The distributions of ancestry proportions were tested against the expected distribution using the non parametric Kolmogorov-Smirnov (KS) test. Correlations between expected and observed estimates of ancestry were ascertained using Spearman’s rank correlation coefficient. All statistical analyses were performed using Minitab v.10.0 software.
Allele frequencies and ancestry information in different markers panels

The balanced marker panel used in the analyses is shown in Table 5.2. The table also shows the allele frequencies and $\delta$ of all pair-wise comparisons. Markers are identified by the SNP consortium (TSC) ID for both panels. The total $\delta$ between EU-WAF is 49.9, total $\delta$ between IA-EU is 56.2, and total $\delta$ between IA-WAF is 57.3. The average $\delta$ between EU-WAF is 0.49, between IA-EU is 0.56 and between IA-WAF is 0.57. Each marker individually is informative for only one axis of ancestry. This represents an ideal situation which should provide the most precise estimates of ancestry.
Table 5.2: Balanced SNP Marker Panel selected from Affymetrix 10k WGA Chip

<table>
<thead>
<tr>
<th>TSCID</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
<th>IA-EU</th>
<th>IA-WAF</th>
<th>EU-WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>889721</td>
<td>0.39</td>
<td>0.04</td>
<td>1</td>
<td>0.35</td>
<td>0.61</td>
<td>0.97</td>
</tr>
<tr>
<td>1554469</td>
<td>1</td>
<td>0.90</td>
<td>0</td>
<td>0.09</td>
<td>1</td>
<td>0.90</td>
</tr>
<tr>
<td>643810</td>
<td>0.17</td>
<td>0.09</td>
<td>0.95</td>
<td>0.07</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>47794</td>
<td>0.03</td>
<td>0.15</td>
<td>1</td>
<td>0.115</td>
<td>0.965</td>
<td>0.85</td>
</tr>
<tr>
<td>640972</td>
<td>0.3</td>
<td>0.16</td>
<td>1</td>
<td>0.14</td>
<td>0.7</td>
<td>0.84</td>
</tr>
<tr>
<td>356065</td>
<td>0.815</td>
<td>0.17</td>
<td>1</td>
<td>0.64</td>
<td>0.19</td>
<td>0.83</td>
</tr>
<tr>
<td>1239268</td>
<td>0.91</td>
<td>1</td>
<td>0.18</td>
<td>0.09</td>
<td>0.72</td>
<td>0.81</td>
</tr>
<tr>
<td>45058</td>
<td>0.24</td>
<td>0.16</td>
<td>0.97</td>
<td>0.08</td>
<td>0.73</td>
<td>0.81</td>
</tr>
<tr>
<td>667224</td>
<td>0.81</td>
<td>0.95</td>
<td>0.15</td>
<td>0.14</td>
<td>0.65</td>
<td>0.80</td>
</tr>
<tr>
<td>1193449</td>
<td>0.21</td>
<td>0.87</td>
<td>0.07</td>
<td>0.66</td>
<td>0.13</td>
<td>0.79</td>
</tr>
<tr>
<td>519349</td>
<td>0</td>
<td>0.12</td>
<td>0.905</td>
<td>0.12</td>
<td>0.905</td>
<td>0.78</td>
</tr>
<tr>
<td>505611</td>
<td>0.29</td>
<td>0.87</td>
<td>0.1</td>
<td>0.58</td>
<td>0.19</td>
<td>0.77</td>
</tr>
<tr>
<td>483243</td>
<td>0.95</td>
<td>0.9</td>
<td>0.12</td>
<td>0.05</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>40555</td>
<td>0</td>
<td>0.76</td>
<td>0</td>
<td>0.76</td>
<td>0</td>
<td>0.76</td>
</tr>
<tr>
<td>272642</td>
<td>0</td>
<td>0.11</td>
<td>0.88</td>
<td>0.11</td>
<td>0.88</td>
<td>0.76</td>
</tr>
<tr>
<td>61423</td>
<td>0.60</td>
<td>0</td>
<td>0.74</td>
<td>0.60</td>
<td>0.14</td>
<td>0.74</td>
</tr>
<tr>
<td>1264567</td>
<td>0</td>
<td>0.77</td>
<td>0.04</td>
<td>0.77</td>
<td>0.04</td>
<td>0.73</td>
</tr>
<tr>
<td>114050</td>
<td>0.97</td>
<td>0.87</td>
<td>0.15</td>
<td>0.10</td>
<td>0.82</td>
<td>0.72</td>
</tr>
<tr>
<td>61718</td>
<td>0.86</td>
<td>0.26</td>
<td>0.97</td>
<td>0.6</td>
<td>0.11</td>
<td>0.71</td>
</tr>
<tr>
<td>521486</td>
<td>0</td>
<td>0.08</td>
<td>0.78</td>
<td>0.08</td>
<td>0.78</td>
<td>0.7</td>
</tr>
<tr>
<td>41409</td>
<td>0.88</td>
<td>0.12</td>
<td>0.82</td>
<td>0.75</td>
<td>0.05</td>
<td>0.7</td>
</tr>
<tr>
<td>511887</td>
<td>0.83</td>
<td>0.16</td>
<td>0.86</td>
<td>0.67</td>
<td>0.02</td>
<td>0.7</td>
</tr>
<tr>
<td>473031</td>
<td>0.02</td>
<td>0.75</td>
<td>0.06</td>
<td>0.73</td>
<td>0.04</td>
<td>0.69</td>
</tr>
<tr>
<td>477639</td>
<td>0.63</td>
<td>0.69</td>
<td>0</td>
<td>0.06</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>450825</td>
<td>0</td>
<td>0.11</td>
<td>0.79</td>
<td>0.11</td>
<td>0.79</td>
<td>0.68</td>
</tr>
<tr>
<td>1011830</td>
<td>0.08</td>
<td>0.18</td>
<td>0.85</td>
<td>0.09</td>
<td>0.77</td>
<td>0.67</td>
</tr>
<tr>
<td>60144</td>
<td>1</td>
<td>0.32</td>
<td>1</td>
<td>0.67</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>58617</td>
<td>0.75</td>
<td>0.67</td>
<td>0</td>
<td>0.08</td>
<td>0.75</td>
<td>0.67</td>
</tr>
<tr>
<td>42798</td>
<td>0</td>
<td>0.67</td>
<td>0</td>
<td>0.67</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>1204940</td>
<td>0</td>
<td>0.13</td>
<td>0.80</td>
<td>0.13</td>
<td>0.80</td>
<td>0.66</td>
</tr>
<tr>
<td>1761621</td>
<td>0.14</td>
<td>0.26</td>
<td>0.93</td>
<td>0.12</td>
<td>0.79</td>
<td>0.66</td>
</tr>
<tr>
<td>37951</td>
<td>0</td>
<td>0.83</td>
<td>0.16</td>
<td>0.83</td>
<td>0.16</td>
<td>0.66</td>
</tr>
<tr>
<td>53987</td>
<td>1</td>
<td>0.18</td>
<td>0.84</td>
<td>0.82</td>
<td>0.15</td>
<td>0.66</td>
</tr>
<tr>
<td>1472708</td>
<td>0.78</td>
<td>0.04</td>
<td>0.7</td>
<td>0.74</td>
<td>0.08</td>
<td>0.65</td>
</tr>
<tr>
<td>TSCID</td>
<td>IA</td>
<td>EU</td>
<td>WAF</td>
<td>IA-EU</td>
<td>IA-WAF</td>
<td>EU-WAF</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>516494</td>
<td>0.98</td>
<td>0.35</td>
<td>1</td>
<td>0.63</td>
<td>0.01</td>
<td>0.65</td>
</tr>
<tr>
<td>43421</td>
<td>0.03</td>
<td>0.12</td>
<td>0.77</td>
<td>0.08</td>
<td>0.73</td>
<td>0.65</td>
</tr>
<tr>
<td>591807</td>
<td>1</td>
<td>0.71</td>
<td>0.07</td>
<td>0.28</td>
<td>0.93</td>
<td>0.64</td>
</tr>
<tr>
<td>912525</td>
<td>1</td>
<td>0.89</td>
<td>0.25</td>
<td>0.10</td>
<td>0.74</td>
<td>0.64</td>
</tr>
<tr>
<td>1002443</td>
<td>0</td>
<td>0.71</td>
<td>0.07</td>
<td>0.71</td>
<td>0.07</td>
<td>0.64</td>
</tr>
<tr>
<td>1250071</td>
<td>0.17</td>
<td>0.35</td>
<td>1</td>
<td>0.18</td>
<td>0.83</td>
<td>0.64</td>
</tr>
<tr>
<td>55197</td>
<td>0</td>
<td>0.11</td>
<td>0.74</td>
<td>0.10</td>
<td>0.74</td>
<td>0.64</td>
</tr>
<tr>
<td>834311</td>
<td>0.94</td>
<td>0.20</td>
<td>0.84</td>
<td>0.73</td>
<td>0.09</td>
<td>0.63</td>
</tr>
<tr>
<td>232054</td>
<td>1</td>
<td>0.88</td>
<td>0.24</td>
<td>0.11</td>
<td>0.75</td>
<td>0.63</td>
</tr>
<tr>
<td>69968</td>
<td>0</td>
<td>0.34</td>
<td>0.97</td>
<td>0.34</td>
<td>0.97</td>
<td>0.63</td>
</tr>
<tr>
<td>980559</td>
<td>0.05</td>
<td>0.15</td>
<td>0.78</td>
<td>0.09</td>
<td>0.73</td>
<td>0.63</td>
</tr>
<tr>
<td>526830</td>
<td>0.21</td>
<td>0.95</td>
<td>0.32</td>
<td>0.75</td>
<td>0.11</td>
<td>0.63</td>
</tr>
<tr>
<td>550050</td>
<td>0.86</td>
<td>0.15</td>
<td>0.785</td>
<td>0.71</td>
<td>0.07</td>
<td>0.63</td>
</tr>
<tr>
<td>151905</td>
<td>1</td>
<td>0.19</td>
<td>0.82</td>
<td>0.80</td>
<td>0.17</td>
<td>0.63</td>
</tr>
<tr>
<td>611274</td>
<td>0</td>
<td>0.09</td>
<td>0.715</td>
<td>0.09</td>
<td>0.71</td>
<td>0.62</td>
</tr>
<tr>
<td>402341</td>
<td>0.96</td>
<td>0.88</td>
<td>0.25</td>
<td>0.08</td>
<td>0.71</td>
<td>0.62</td>
</tr>
<tr>
<td>764714</td>
<td>0.97</td>
<td>0.25</td>
<td>0.88</td>
<td>0.71</td>
<td>0.09</td>
<td>0.62</td>
</tr>
<tr>
<td>345031</td>
<td>0.09</td>
<td>0.77</td>
<td>0.15</td>
<td>0.68</td>
<td>0.06</td>
<td>0.62</td>
</tr>
<tr>
<td>48343</td>
<td>1</td>
<td>0.33</td>
<td>0.95</td>
<td>0.66</td>
<td>0.04</td>
<td>0.62</td>
</tr>
<tr>
<td>575401</td>
<td>1</td>
<td>0.33</td>
<td>0.94</td>
<td>0.67</td>
<td>0.05</td>
<td>0.61</td>
</tr>
<tr>
<td>16204</td>
<td>0.26</td>
<td>0.37</td>
<td>0.99</td>
<td>0.11</td>
<td>0.72</td>
<td>0.61</td>
</tr>
<tr>
<td>543384</td>
<td>0.06</td>
<td>0.17</td>
<td>0.785</td>
<td>0.11</td>
<td>0.725</td>
<td>0.62</td>
</tr>
<tr>
<td>145717</td>
<td>0.05</td>
<td>0.79</td>
<td>0.19</td>
<td>0.74</td>
<td>0.135</td>
<td>0.60</td>
</tr>
<tr>
<td>299899</td>
<td>0.84</td>
<td>0</td>
<td>0.6</td>
<td>0.84</td>
<td>0.24</td>
<td>0.6</td>
</tr>
<tr>
<td>575402</td>
<td>0</td>
<td>0.70</td>
<td>0.10</td>
<td>0.70</td>
<td>0.10</td>
<td>0.6</td>
</tr>
<tr>
<td>1270991</td>
<td>0.85</td>
<td>0.29</td>
<td>0.89</td>
<td>0.55</td>
<td>0.04</td>
<td>0.6</td>
</tr>
<tr>
<td>1133193</td>
<td>1</td>
<td>0.89</td>
<td>0.29</td>
<td>0.10</td>
<td>0.70</td>
<td>0.59</td>
</tr>
<tr>
<td>481419</td>
<td>0.06</td>
<td>0.88</td>
<td>0.28</td>
<td>0.82</td>
<td>0.23</td>
<td>0.59</td>
</tr>
<tr>
<td>54875</td>
<td>0.90</td>
<td>0.77</td>
<td>0.17</td>
<td>0.13</td>
<td>0.73</td>
<td>0.59</td>
</tr>
<tr>
<td>602736</td>
<td>0.06</td>
<td>0.78</td>
<td>0.18</td>
<td>0.715</td>
<td>0.12</td>
<td>0.59</td>
</tr>
<tr>
<td>942395</td>
<td>0</td>
<td>0.68</td>
<td>0.09</td>
<td>0.68</td>
<td>0.09</td>
<td>0.59</td>
</tr>
<tr>
<td>363686</td>
<td>0.12</td>
<td>0.81</td>
<td>0.22</td>
<td>0.69</td>
<td>0.10</td>
<td>0.58</td>
</tr>
<tr>
<td>60526</td>
<td>1</td>
<td>0.64</td>
<td>0.06</td>
<td>0.35</td>
<td>0.94</td>
<td>0.58</td>
</tr>
<tr>
<td>539848</td>
<td>0.05</td>
<td>0.86</td>
<td>0.31</td>
<td>0.81</td>
<td>0.25</td>
<td>0.55</td>
</tr>
<tr>
<td>171306</td>
<td>1</td>
<td>0.03</td>
<td>0.305</td>
<td>0.97</td>
<td>0.69</td>
<td>0.27</td>
</tr>
</tbody>
</table>
Table 5.2 Continued

<table>
<thead>
<tr>
<th>TSCID</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
<th>IA-EU δ</th>
<th>IA-WAF δ</th>
<th>EU-WAF δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1274868</td>
<td>0</td>
<td>0.97</td>
<td>0.73</td>
<td>0.97</td>
<td>0.73</td>
<td>0.24</td>
</tr>
<tr>
<td>406303</td>
<td>0.91</td>
<td>0.21</td>
<td>0</td>
<td>0.7</td>
<td>0.90</td>
<td>0.21</td>
</tr>
<tr>
<td>149641</td>
<td>1</td>
<td>0.19</td>
<td>0.39</td>
<td>0.81</td>
<td>0.60</td>
<td>0.21</td>
</tr>
<tr>
<td>759465</td>
<td>0</td>
<td>0.83</td>
<td>0.63</td>
<td>0.83</td>
<td>0.63</td>
<td>0.20</td>
</tr>
<tr>
<td>261103</td>
<td>1</td>
<td>0.21</td>
<td>0.38</td>
<td>0.79</td>
<td>0.61</td>
<td>0.18</td>
</tr>
<tr>
<td>941538</td>
<td>0.12</td>
<td>0.95</td>
<td>0.78</td>
<td>0.83</td>
<td>0.66</td>
<td>0.17</td>
</tr>
<tr>
<td>922911</td>
<td>1</td>
<td>0.22</td>
<td>0.07</td>
<td>0.77</td>
<td>0.92</td>
<td>0.15</td>
</tr>
<tr>
<td>46389</td>
<td>0.88</td>
<td>0.14</td>
<td>0</td>
<td>0.73</td>
<td>0.88</td>
<td>0.14</td>
</tr>
<tr>
<td>1551899</td>
<td>0.07</td>
<td>0.77</td>
<td>0.91</td>
<td>0.70</td>
<td>0.83</td>
<td>0.13</td>
</tr>
<tr>
<td>268195</td>
<td>0.92</td>
<td>0.13</td>
<td>0</td>
<td>0.79</td>
<td>0.92</td>
<td>0.13</td>
</tr>
<tr>
<td>75980</td>
<td>0.06</td>
<td>0.93</td>
<td>0.81</td>
<td>0.86</td>
<td>0.73</td>
<td>0.13</td>
</tr>
<tr>
<td>53429</td>
<td>0.92</td>
<td>0.14</td>
<td>0.02</td>
<td>0.77</td>
<td>0.9</td>
<td>0.12</td>
</tr>
<tr>
<td>268196</td>
<td>0.91</td>
<td>0.11</td>
<td>0</td>
<td>0.80</td>
<td>0.91</td>
<td>0.11</td>
</tr>
<tr>
<td>602822</td>
<td>0.13</td>
<td>0.89</td>
<td>1</td>
<td>0.75</td>
<td>0.86</td>
<td>0.11</td>
</tr>
<tr>
<td>40569</td>
<td>1</td>
<td>0.07</td>
<td>0.17</td>
<td>0.92</td>
<td>0.82</td>
<td>0.10</td>
</tr>
<tr>
<td>549661</td>
<td>1</td>
<td>0.19</td>
<td>0.10</td>
<td>0.80</td>
<td>0.89</td>
<td>0.09</td>
</tr>
<tr>
<td>649738</td>
<td>0.92</td>
<td>0.06</td>
<td>0.15</td>
<td>0.86</td>
<td>0.77</td>
<td>0.09</td>
</tr>
<tr>
<td>376342</td>
<td>0.08</td>
<td>1</td>
<td>0.92</td>
<td>0.91</td>
<td>0.83</td>
<td>0.07</td>
</tr>
<tr>
<td>43628</td>
<td>0.04</td>
<td>0.25</td>
<td>0.85</td>
<td>0.74</td>
<td>0.81</td>
<td>0.07</td>
</tr>
<tr>
<td>53865</td>
<td>0.14</td>
<td>0.89</td>
<td>0.82</td>
<td>0.74</td>
<td>0.67</td>
<td>0.07</td>
</tr>
<tr>
<td>916690</td>
<td>0</td>
<td>0.79</td>
<td>0.86</td>
<td>0.79</td>
<td>0.86</td>
<td>0.06</td>
</tr>
<tr>
<td>65289</td>
<td>0.79</td>
<td>0.05</td>
<td>0</td>
<td>0.73</td>
<td>0.79</td>
<td>0.05</td>
</tr>
<tr>
<td>1762383</td>
<td>0.02</td>
<td>0.75</td>
<td>0.80</td>
<td>0.72</td>
<td>0.78</td>
<td>0.05</td>
</tr>
<tr>
<td>800273</td>
<td>0</td>
<td>0.81</td>
<td>0.85</td>
<td>0.81</td>
<td>0.85</td>
<td>0.04</td>
</tr>
<tr>
<td>256761</td>
<td>0.95</td>
<td>0.10</td>
<td>0.14</td>
<td>0.84</td>
<td>0.80</td>
<td>0.04</td>
</tr>
<tr>
<td>527086</td>
<td>1</td>
<td>0.22</td>
<td>0.26</td>
<td>0.77</td>
<td>0.73</td>
<td>0.03</td>
</tr>
<tr>
<td>51052</td>
<td>0.14</td>
<td>0.89</td>
<td>0.92</td>
<td>0.75</td>
<td>0.78</td>
<td>0.03</td>
</tr>
<tr>
<td>721867</td>
<td>0.93</td>
<td>0.12</td>
<td>0.08</td>
<td>0.81</td>
<td>0.84</td>
<td>0.03</td>
</tr>
<tr>
<td>26430</td>
<td>0</td>
<td>0.74</td>
<td>0.71</td>
<td>0.74</td>
<td>0.71</td>
<td>0.03</td>
</tr>
<tr>
<td>855434</td>
<td>0.8</td>
<td>0.05</td>
<td>0.02</td>
<td>0.74</td>
<td>0.77</td>
<td>0.03</td>
</tr>
<tr>
<td>927612</td>
<td>0.05</td>
<td>0.89</td>
<td>0.87</td>
<td>0.84</td>
<td>0.82</td>
<td>0.02</td>
</tr>
<tr>
<td>52786</td>
<td>0.93</td>
<td>0.09</td>
<td>0.10</td>
<td>0.83</td>
<td>0.83</td>
<td>0.00</td>
</tr>
<tr>
<td>1258006</td>
<td>0.08</td>
<td>1</td>
<td>1</td>
<td>0.92</td>
<td>0.92</td>
<td>0</td>
</tr>
</tbody>
</table>

| Total δ  | 56.16 | 57.27 | 49.94 |
| Average δ | 0.55  | 0.56  | 0.49  |
The unbalanced panel used for the analyses is shown in Table 5.3. The markers were selected based on high EU-WAF $\delta$ with no attention to the IA-EU or IA-WAF $\delta$ values, since the intention was to distinguish between West African and European populations only. Total EU-WAF $\delta$ in this panel is 79.8, with an average $\delta$ value of 0.78. Corresponding $\delta$ values for IA-WAF is 66.5 with average values of 0.65 and $\delta$ for IA-EU is 17.7 with average $\delta$ value of 0.17. This panel represents another extreme, but very likely scenario when markers are selected based on major populations.
<table>
<thead>
<tr>
<th>TSCID</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
<th>IA-EU δ</th>
<th>IA-WAF δ</th>
<th>EU-WAF δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>500306</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>889721</td>
<td>0.36</td>
<td>0.02</td>
<td>1</td>
<td>0.33</td>
<td>0.97</td>
<td>0.63</td>
</tr>
<tr>
<td>53441</td>
<td>0.16</td>
<td>0.02</td>
<td>0.95</td>
<td>0.13</td>
<td>0.93</td>
<td>0.79</td>
</tr>
<tr>
<td>755351</td>
<td>0.87</td>
<td>0.95</td>
<td>0.02</td>
<td>0.07</td>
<td>0.92</td>
<td>0.84</td>
</tr>
<tr>
<td>55196</td>
<td>0.13</td>
<td>0.07</td>
<td>0.97</td>
<td>0.06</td>
<td>0.90</td>
<td>0.83</td>
</tr>
<tr>
<td>47849</td>
<td>0.52</td>
<td>0.07</td>
<td>0.97</td>
<td>0.45</td>
<td>0.90</td>
<td>0.45</td>
</tr>
<tr>
<td>702103</td>
<td>0.92</td>
<td>0.9</td>
<td>0</td>
<td>0.02</td>
<td>0.9</td>
<td>0.92</td>
</tr>
<tr>
<td>863187</td>
<td>1</td>
<td>0.89</td>
<td>0.10</td>
<td>1</td>
<td>0.89</td>
<td>1</td>
</tr>
<tr>
<td>39432</td>
<td>1</td>
<td>1</td>
<td>0.11</td>
<td>0</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>643810</td>
<td>0.16</td>
<td>0.07</td>
<td>0.95</td>
<td>0.08</td>
<td>0.87</td>
<td>0.79</td>
</tr>
<tr>
<td>253130</td>
<td>0.87</td>
<td>0.86</td>
<td>0</td>
<td>0.01</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td>899340</td>
<td>0.15</td>
<td>0.14</td>
<td>1</td>
<td>0.00</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>1336441</td>
<td>0.69</td>
<td>0.85</td>
<td>0</td>
<td>0.16</td>
<td>0.85</td>
<td>0.69</td>
</tr>
<tr>
<td>57627</td>
<td>0.81</td>
<td>0.85</td>
<td>0</td>
<td>0.03</td>
<td>0.85</td>
<td>0.81</td>
</tr>
<tr>
<td>47794</td>
<td>0.08</td>
<td>0.15</td>
<td>1</td>
<td>0.07</td>
<td>0.84</td>
<td>0.91</td>
</tr>
<tr>
<td>794164</td>
<td>0.97</td>
<td>1</td>
<td>0.15</td>
<td>0.02</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>52635</td>
<td>0.86</td>
<td>0.92</td>
<td>0.09</td>
<td>0.05</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>1554469</td>
<td>0.98</td>
<td>0.9</td>
<td>0.06</td>
<td>0.08</td>
<td>0.83</td>
<td>0.91</td>
</tr>
<tr>
<td>380878</td>
<td>0.91</td>
<td>0.9</td>
<td>0.06</td>
<td>0.01</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td>42830</td>
<td>0.67</td>
<td>0.92</td>
<td>0.09</td>
<td>0.24</td>
<td>0.83</td>
<td>0.58</td>
</tr>
<tr>
<td>45058</td>
<td>0.31</td>
<td>0.15</td>
<td>0.97</td>
<td>0.16</td>
<td>0.82</td>
<td>0.66</td>
</tr>
<tr>
<td>57690</td>
<td>0.48</td>
<td>0.15</td>
<td>0.97</td>
<td>0.33</td>
<td>0.82</td>
<td>0.48</td>
</tr>
<tr>
<td>272275</td>
<td>0.28</td>
<td>0.05</td>
<td>0.87</td>
<td>0.22</td>
<td>0.81</td>
<td>0.59</td>
</tr>
<tr>
<td>1444703</td>
<td>0.51</td>
<td>0.07</td>
<td>0.89</td>
<td>0.43</td>
<td>0.81</td>
<td>0.38</td>
</tr>
<tr>
<td>851662</td>
<td>0.98</td>
<td>0.86</td>
<td>0.04</td>
<td>0.12</td>
<td>0.82</td>
<td>0.94</td>
</tr>
<tr>
<td>264710</td>
<td>0.5</td>
<td>0.81</td>
<td>0</td>
<td>0.31</td>
<td>0.82</td>
<td>0.5</td>
</tr>
<tr>
<td>1143600</td>
<td>0.25</td>
<td>0.09</td>
<td>0.9</td>
<td>0.16</td>
<td>0.82</td>
<td>0.65</td>
</tr>
<tr>
<td>57526</td>
<td>0.07</td>
<td>0.05</td>
<td>0.86</td>
<td>0.02</td>
<td>0.81</td>
<td>0.78</td>
</tr>
<tr>
<td>991109</td>
<td>0.15</td>
<td>0.18</td>
<td>1</td>
<td>0.03</td>
<td>0.81</td>
<td>0.85</td>
</tr>
<tr>
<td>48235</td>
<td>1</td>
<td>1</td>
<td>0.18</td>
<td>0</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>64519</td>
<td>0.42</td>
<td>0.93</td>
<td>0.11</td>
<td>0.5</td>
<td>0.81</td>
<td>0.31</td>
</tr>
<tr>
<td>519349</td>
<td>0.02</td>
<td>0.1</td>
<td>0.9</td>
<td>0.07</td>
<td>0.81</td>
<td>0.88</td>
</tr>
<tr>
<td>163515</td>
<td>1</td>
<td>0.81</td>
<td>0</td>
<td>0.19</td>
<td>0.81</td>
<td>1</td>
</tr>
<tr>
<td>515655</td>
<td>0.59</td>
<td>0.85</td>
<td>0.04</td>
<td>0.26</td>
<td>0.81</td>
<td>0.54</td>
</tr>
<tr>
<td>TSCID</td>
<td>IA</td>
<td>EU</td>
<td>WAF</td>
<td>IA-EU δ</td>
<td>IA-WAF δ</td>
<td>EU-WAF δ</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>---------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>531007</td>
<td>0.31</td>
<td>0.17</td>
<td>0.97</td>
<td>0.13</td>
<td>0.80</td>
<td>0.66</td>
</tr>
<tr>
<td>1530248</td>
<td>0.61</td>
<td>0.9</td>
<td>0.1</td>
<td>0.28</td>
<td>0.8</td>
<td>0.51</td>
</tr>
<tr>
<td>59030</td>
<td>0.54</td>
<td>0.92</td>
<td>0.13</td>
<td>0.38</td>
<td>0.79</td>
<td>0.40</td>
</tr>
<tr>
<td>533890</td>
<td>0.02</td>
<td>0.15</td>
<td>0.94</td>
<td>0.12</td>
<td>0.79</td>
<td>0.91</td>
</tr>
<tr>
<td>428601</td>
<td>0.17</td>
<td>0.21</td>
<td>1</td>
<td>0.03</td>
<td>0.78</td>
<td>0.82</td>
</tr>
<tr>
<td>954841</td>
<td>0.2</td>
<td>0.12</td>
<td>0.91</td>
<td>0.07</td>
<td>0.78</td>
<td>0.70</td>
</tr>
<tr>
<td>856010</td>
<td>0.53</td>
<td>0.87</td>
<td>0.09</td>
<td>0.33</td>
<td>0.78</td>
<td>0.44</td>
</tr>
<tr>
<td>270847</td>
<td>0.07</td>
<td>0.22</td>
<td>1</td>
<td>0.14</td>
<td>0.77</td>
<td>0.92</td>
</tr>
<tr>
<td>875649</td>
<td>1</td>
<td>1</td>
<td>0.23</td>
<td>0</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>64860</td>
<td>0.02</td>
<td>0</td>
<td>0.77</td>
<td>0.02</td>
<td>0.77</td>
<td>0.74</td>
</tr>
<tr>
<td>40555</td>
<td>0.04</td>
<td>0.76</td>
<td>0</td>
<td>0.72</td>
<td>0.76</td>
<td>0.04</td>
</tr>
<tr>
<td>640972</td>
<td>0.22</td>
<td>0.18</td>
<td>0.95</td>
<td>0.04</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>326382</td>
<td>1</td>
<td>0.95</td>
<td>0.18</td>
<td>0.05</td>
<td>0.76</td>
<td>0.81</td>
</tr>
<tr>
<td>50865</td>
<td>0.80</td>
<td>0.23</td>
<td>1</td>
<td>0.56</td>
<td>0.76</td>
<td>0.19</td>
</tr>
<tr>
<td>61423</td>
<td>0.48</td>
<td>0</td>
<td>0.76</td>
<td>0.48</td>
<td>0.76</td>
<td>0.27</td>
</tr>
<tr>
<td>64440</td>
<td>1</td>
<td>0.87</td>
<td>0.11</td>
<td>0.12</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>241894</td>
<td>0.23</td>
<td>0.20</td>
<td>0.96</td>
<td>0.03</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>540297</td>
<td>0.25</td>
<td>0.15</td>
<td>0.91</td>
<td>0.1</td>
<td>0.75</td>
<td>0.65</td>
</tr>
<tr>
<td>62321</td>
<td>0.56</td>
<td>0.15</td>
<td>0.91</td>
<td>0.41</td>
<td>0.75</td>
<td>0.34</td>
</tr>
<tr>
<td>714589</td>
<td>0.82</td>
<td>0.82</td>
<td>0.06</td>
<td>0</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>667224</td>
<td>0.68</td>
<td>0.93</td>
<td>0.18</td>
<td>0.25</td>
<td>0.75</td>
<td>0.50</td>
</tr>
<tr>
<td>1291477</td>
<td>0.20</td>
<td>0.77</td>
<td>0.02</td>
<td>0.57</td>
<td>0.75</td>
<td>0.17</td>
</tr>
<tr>
<td>58237</td>
<td>0.28</td>
<td>0.13</td>
<td>0.88</td>
<td>0.15</td>
<td>0.75</td>
<td>0.59</td>
</tr>
<tr>
<td>587842</td>
<td>0.21</td>
<td>0.079</td>
<td>0.93</td>
<td>0.03</td>
<td>0.75</td>
<td>0.71</td>
</tr>
<tr>
<td>603488</td>
<td>0.21</td>
<td>0.75</td>
<td>0.02</td>
<td>0.56</td>
<td>0.75</td>
<td>0.18</td>
</tr>
<tr>
<td>988681</td>
<td>0.98</td>
<td>0.75</td>
<td>0.12</td>
<td>0.11</td>
<td>0.75</td>
<td>0.86</td>
</tr>
<tr>
<td>516492</td>
<td>1</td>
<td>1</td>
<td>0.25</td>
<td>0</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>840999</td>
<td>0.75</td>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>54280</td>
<td>0.97</td>
<td>1</td>
<td>0.25</td>
<td>0.02</td>
<td>0.75</td>
<td>0.72</td>
</tr>
<tr>
<td>39147</td>
<td>0.82</td>
<td>0.95</td>
<td>0.2</td>
<td>0.12</td>
<td>0.75</td>
<td>0.62</td>
</tr>
<tr>
<td>42022</td>
<td>0.27</td>
<td>0.12</td>
<td>0.85</td>
<td>0.15</td>
<td>0.75</td>
<td>0.6</td>
</tr>
<tr>
<td>61424</td>
<td>0.49</td>
<td>1</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
<td>0.24</td>
</tr>
<tr>
<td>46225</td>
<td>0.23</td>
<td>0.75</td>
<td>0</td>
<td>0.51</td>
<td>0.75</td>
<td>0.23</td>
</tr>
<tr>
<td>502179</td>
<td>0.12</td>
<td>0.025</td>
<td>0.77</td>
<td>0.1</td>
<td>0.74</td>
<td>0.64</td>
</tr>
<tr>
<td>TSCID</td>
<td>IA</td>
<td>EU</td>
<td>WAF</td>
<td>IA-EU</td>
<td>IA-WAF</td>
<td>EU-WAF</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>485139</td>
<td>0.1</td>
<td>0.05</td>
<td>0.79</td>
<td>0.05</td>
<td>0.74</td>
<td>0.69</td>
</tr>
<tr>
<td>507734</td>
<td>0.69</td>
<td>0.81</td>
<td>0.06</td>
<td>0.12</td>
<td>0.74</td>
<td>0.62</td>
</tr>
<tr>
<td>392655</td>
<td>0.8</td>
<td>0.95</td>
<td>0.20</td>
<td>0.15</td>
<td>0.74</td>
<td>0.59</td>
</tr>
<tr>
<td>58017</td>
<td>0</td>
<td>0.07</td>
<td>0.81</td>
<td>0.07</td>
<td>0.74</td>
<td>0.81</td>
</tr>
<tr>
<td>483243</td>
<td>0.91</td>
<td>0.87</td>
<td>0.13</td>
<td>0.04</td>
<td>0.74</td>
<td>0.78</td>
</tr>
<tr>
<td>59404</td>
<td>0.91</td>
<td>0.9</td>
<td>0.15</td>
<td>0.01</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>10190</td>
<td>0.72</td>
<td>0.94</td>
<td>0.20</td>
<td>0.22</td>
<td>0.74</td>
<td>0.51</td>
</tr>
<tr>
<td>48138</td>
<td>0.22</td>
<td>0.76</td>
<td>0.02</td>
<td>0.53</td>
<td>0.74</td>
<td>0.20</td>
</tr>
<tr>
<td>844539</td>
<td>0.06</td>
<td>0.02</td>
<td>0.76</td>
<td>0.03</td>
<td>0.73</td>
<td>0.69</td>
</tr>
<tr>
<td>662511</td>
<td>0.10</td>
<td>0.07</td>
<td>0.81</td>
<td>0.03</td>
<td>0.73</td>
<td>0.70</td>
</tr>
<tr>
<td>675833</td>
<td>0.04</td>
<td>0.03</td>
<td>0.77</td>
<td>0.00</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>505611</td>
<td>0.18</td>
<td>0.85</td>
<td>0.11</td>
<td>0.66</td>
<td>0.73</td>
<td>0.07</td>
</tr>
<tr>
<td>51129</td>
<td>0.06</td>
<td>0.17</td>
<td>0.90</td>
<td>0.11</td>
<td>0.73</td>
<td>0.84</td>
</tr>
<tr>
<td>517161</td>
<td>0.68</td>
<td>0.91</td>
<td>0.18</td>
<td>0.23</td>
<td>0.73</td>
<td>0.50</td>
</tr>
<tr>
<td>39291</td>
<td>0.96</td>
<td>0.93</td>
<td>0.20</td>
<td>0.02</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>49480</td>
<td>0.85</td>
<td>0.83</td>
<td>0.1</td>
<td>0.02</td>
<td>0.73</td>
<td>0.75</td>
</tr>
<tr>
<td>59671</td>
<td>0.60</td>
<td>0.22</td>
<td>0.95</td>
<td>0.38</td>
<td>0.73</td>
<td>0.35</td>
</tr>
<tr>
<td>717625</td>
<td>0.13</td>
<td>0.12</td>
<td>0.85</td>
<td>0.01</td>
<td>0.73</td>
<td>0.71</td>
</tr>
<tr>
<td>56876</td>
<td>0.22</td>
<td>0.2</td>
<td>0.93</td>
<td>0.02</td>
<td>0.73</td>
<td>0.70</td>
</tr>
<tr>
<td>58816</td>
<td>0.62</td>
<td>0.2</td>
<td>0.93</td>
<td>0.42</td>
<td>0.73</td>
<td>0.30</td>
</tr>
<tr>
<td>1426297</td>
<td>0.04</td>
<td>0.13</td>
<td>0.86</td>
<td>0.09</td>
<td>0.73</td>
<td>0.82</td>
</tr>
<tr>
<td>1243732</td>
<td>0.76</td>
<td>0.77</td>
<td>0.04</td>
<td>0.01</td>
<td>0.73</td>
<td>0.71</td>
</tr>
<tr>
<td>494968</td>
<td>0</td>
<td>0</td>
<td>0.72</td>
<td>0</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>41752</td>
<td>0.58</td>
<td>0.25</td>
<td>0.97</td>
<td>0.33</td>
<td>0.72</td>
<td>0.38</td>
</tr>
<tr>
<td>450825</td>
<td>0.05</td>
<td>0.1</td>
<td>0.82</td>
<td>0.04</td>
<td>0.72</td>
<td>0.76</td>
</tr>
<tr>
<td>55903</td>
<td>0.97</td>
<td>0.84</td>
<td>0.11</td>
<td>0.13</td>
<td>0.72</td>
<td>0.85</td>
</tr>
<tr>
<td>169415</td>
<td>0.04</td>
<td>0.02</td>
<td>0.75</td>
<td>0.01</td>
<td>0.72</td>
<td>0.70</td>
</tr>
<tr>
<td>506462</td>
<td>0.11</td>
<td>0.05</td>
<td>0.77</td>
<td>0.06</td>
<td>0.72</td>
<td>0.65</td>
</tr>
<tr>
<td>65099</td>
<td>0.22</td>
<td>0.27</td>
<td>1</td>
<td>0.05</td>
<td>0.72</td>
<td>0.77</td>
</tr>
<tr>
<td>596416</td>
<td>0.9</td>
<td>0.92</td>
<td>0.20</td>
<td>0.02</td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td>473031</td>
<td>0.02</td>
<td>0.78</td>
<td>0.07</td>
<td>0.76</td>
<td>0.71</td>
<td>0.05</td>
</tr>
<tr>
<td>441868</td>
<td>0.51</td>
<td>0.73</td>
<td>0.02</td>
<td>0.22</td>
<td>0.71</td>
<td>0.48</td>
</tr>
<tr>
<td>521486</td>
<td>0</td>
<td>0.1</td>
<td>0.81</td>
<td>0.1</td>
<td>0.71</td>
<td>0.81</td>
</tr>
<tr>
<td>44825</td>
<td>0.41</td>
<td>0.95</td>
<td>0.23</td>
<td>0.53</td>
<td>0.71</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Effect of segregation

Table 5.1 shows the distribution of genealogical ancestry in the sample. Shown are 17 groups and within each group all individuals (number of individuals in each group shown in first column) have the same genealogical ancestry (the expected ancestral level based on family structure – See Chapter 1 for details). All simulations were started with 512 ancestors 9 generations in the past. Order of ancestral individuals in the first generation was randomly assigned and no effect of the order of individuals was observed on the final estimates of ancestry (data not shown). Total simulated sample consists of 603 individuals with varying ancestry proportions.

Table 5.4 shows the results for all analyses using the balanced panel of markers and Table 5.5 shows the results for all analyses using the unbalanced panel of markers. As seen in table 5.4, mean segregation ancestry in the sample is 0.69 for WAF, 0.25 for EU and 0.05 for IA. This estimate represents the weighted mean for the entire sample. These estimates are used as the expected estimates for comparing against estimates obtained with each program/option. Corresponding means for expected segregation ancestry using the unbalanced marker panel is shown in Table 5 and is 0.701 WAF, 0.244 EU and 0.055 IA.
Table 5.4: Results for Balanced Panel of markers: Mean (Variance) for each ancestry axis, MSE, Correlations between expected and observed estimates of ancestry and Results of t-test and KS tests

<table>
<thead>
<tr>
<th>SIM</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
<th>MSE</th>
<th>Correlations</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGR</td>
<td>0.051 (0.002)</td>
<td>0.25 (0.022)</td>
<td>0.699 (0.024)</td>
<td>0.006</td>
<td><strong>0.73; 0.95; 0.96</strong></td>
<td>0.06; 0.04; 0.05</td>
</tr>
<tr>
<td>MLE-B</td>
<td>0.052 (0.003)</td>
<td>0.25 (0.024)</td>
<td>0.697 (0.026)</td>
<td>0.009</td>
<td><strong>0.53; 0.94; 0.95</strong></td>
<td><strong>0.38; 0.15; 0.08</strong></td>
</tr>
<tr>
<td>NP-I-B</td>
<td>0.029 (0.001)*</td>
<td>0.29 (0.022)*</td>
<td>0.678 (0.024)*</td>
<td>0.005</td>
<td><strong>0.74; 0.94; 0.95</strong></td>
<td><strong>0.21; 0.09; 0.08</strong></td>
</tr>
<tr>
<td>NP-S-B</td>
<td>0.32 (0.043)*</td>
<td>0.307 (0.026)*</td>
<td>0.375 (0.06)*</td>
<td>0.0278</td>
<td><strong>0.11; 0.17; 0.92</strong></td>
<td><strong>0.72; 0.13; 0.67</strong></td>
</tr>
<tr>
<td>F-I-B</td>
<td>0.041 (0.001)*</td>
<td>0.258 (0.018)</td>
<td>0.701 (0.019)</td>
<td>0.005</td>
<td><strong>0.74; 0.95; 0.95</strong></td>
<td><strong>0.2; 0.09; 0.08</strong></td>
</tr>
<tr>
<td>F-S-B</td>
<td>0.058 (0.002)*</td>
<td>0.25 (0.023)</td>
<td>0.692 (0.025)</td>
<td>0.005</td>
<td><strong>0.74; 0.95; 0.95</strong></td>
<td><strong>0.15; 0.08; 0.08</strong></td>
</tr>
<tr>
<td>NF-I-B</td>
<td>0.041 (0.001)*</td>
<td>0.26 (0.019)</td>
<td>0.701 (0.02)</td>
<td>0.005</td>
<td><strong>0.74; 0.95; 0.95</strong></td>
<td><strong>0.2; 0.09; 0.08</strong></td>
</tr>
<tr>
<td>NF-S-B</td>
<td>0.058 (0.002)*</td>
<td>0.25 (0.023)</td>
<td>0.691 (0.025)</td>
<td>0.005</td>
<td><strong>0.74; 0.95; 0.95</strong></td>
<td><strong>0.15; 0.08; 0.09</strong></td>
</tr>
</tbody>
</table>

SGR: Segregation ancestry; MLE: Estimates obtained with ML;
NP-I-B: STRUCTURE model with no parentals and independent alpha parameter
NP-S-B: STRUCTURE model with no parentals and same alpha parameter
F-I-B: STRUCTURE model with parentals and specifying “Usepopinfo” and “Popflag” and independent alpha parameter
F-S-B: STRUCTURE model with parentals and specifying “Usepopinfo” and “Popflag” and same alpha parameter
NF-I-B: STRUCTURE model with parentals specifying “Usepopinfo” and independent alpha parameter
NF-S-B: STRUCTURE model with parentals specifying “Usepopinfo” and same alpha parameter
IA: Mean (Variance) of Indigenous American ancestry; EU: Mean (Variance) of European ancestry; WAF: Mean (Variance) of West African ancestry
Correlations: Showing Spearman’s correlation coefficient for each ancestry axis in the order IA, EU and WAF. All correlations were significant (P<0.0001) except one which is not in bold.
KS: Results of Kolmogorov-Smirnov test showing D value. Significant D values (P<0.01 after Bonferroni correction) are shown in bold, indicating significant departure from expected distributions
* Significant t tests comparing corresponding segregation ancestry to estimated ancestry from each axis after Bonferroni correction
Table 5.5: Results for Unbalanced Panel of markers: Mean (Variance) for each ancestry axis, MSE, Correlations between expected and observed estimates of ancestry and Results of t-test and KS tests

<table>
<thead>
<tr>
<th>SIM</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
<th>MSE</th>
<th>Correlations</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGR</td>
<td>0.055 (0.002)</td>
<td>0.244 (0.02)</td>
<td>0.701 (0.025)</td>
<td></td>
<td>IA; EU; WAF</td>
<td></td>
</tr>
<tr>
<td>MLE-UB</td>
<td>0.087 (0.01)*</td>
<td>0.217 (0.024)*</td>
<td>0.695 (0.026)</td>
<td>0.019</td>
<td>0.36; 0.82; 0.97</td>
<td>0.27; 0.13; 0.04</td>
</tr>
<tr>
<td>NP-I-UB</td>
<td>0.028 (0.003)*</td>
<td>0.322 (0.028)*</td>
<td>0.649 (0.031)*</td>
<td>0.015</td>
<td>0.29; 0.94; 0.98</td>
<td>0.44; 0.26; 0.18</td>
</tr>
<tr>
<td>NP-S-UB</td>
<td>0.285 (0.042)*</td>
<td>0.325 (0.032)*</td>
<td>0.389 (0.067)*</td>
<td>0.028</td>
<td>0.45; 0.31; 0.92</td>
<td>0.67; 0.21; 0.62</td>
</tr>
<tr>
<td>F-I-UB</td>
<td>0.046 (0.001)*</td>
<td>0.251 (0.018)</td>
<td>0.702 (0.022)</td>
<td>0.005</td>
<td>0.49; 0.94; 0.98</td>
<td>0.18; 0.07; 0.07</td>
</tr>
<tr>
<td>F-S-UB</td>
<td>0.114 (0.006)*</td>
<td>0.198 (0.019)*</td>
<td>0.687 (0.026)</td>
<td>0.016</td>
<td>0.47; 0.88; 0.98</td>
<td>0.37; 0.21; 0.08</td>
</tr>
<tr>
<td>NF-I-UB</td>
<td>0.109 (0.02) *</td>
<td>0.34 (0.019) *</td>
<td>0.55 (0.102) *</td>
<td>0.448</td>
<td>0.45; -0.73; -0.54</td>
<td>0.19; 0.29; 0.37</td>
</tr>
<tr>
<td>NF-S-UB</td>
<td>0.088 (0.008) *</td>
<td>0.213 (0.02) *</td>
<td>0.069 (0.02) *</td>
<td>0.017</td>
<td>0.39; 0.85; 0.97</td>
<td>0.18; 0.14; 0.05</td>
</tr>
</tbody>
</table>

SGR: Segregation ancestry; MLE-UB: Estimates obtained with ML; NP-I-UB: STRUCTURE model with no parentals and independent alpha parameter; NP-S-UB: STRUCTURE model with no parentals and same alpha parameter; F-I-UB: STRUCTURE model with parentals and specifying “Usepopinfo” and “Popflag” and independent alpha parameter; F-S-UB: STRUCTURE model with parentals and specifying “Usepopinfo” and “Popflag” and same alpha parameter; NF-I-UB: STRUCTURE model with parentals specifying “Usepopinfo” and independent alpha parameter; NF-S-UB: STRUCTURE model with parentals specifying “Usepopinfo” and same alpha parameter; IA: Mean (Variance) of Indigenous American ancestry; EU: Mean (Variance) of European ancestry; WAF: Mean (Variance) of West African ancestry; MSE: Mean Square error for each option; Correlations: Showing Spearman’s correlation coefficient for each ancestry axis in the order IA, EU and WAF. All correlations were significant (P<0.0001) except one which is shown in bold; KS: Results of Kolmogorov-Smirnov test showing D value. Significant D values (P<0.01 after Bonferroni correction) are shown in bold, indicating significant departure from expected distributions; * Significant t tests comparing segregation ancestry to estimated ancestry from each axis after Bonferroni correction.
The plot labeled SGR-B in Figure 5.1 shows the distribution of segregation ancestry proportions in the sample using the balanced marker panel. Plot SGR-UB in Figure 5.2 shows the distribution of expected segregation ancestry estimates in sample using the unbalanced marker panel.

Figure 5.1: Triangle plots of ancestry estimates in simulated samples using the balanced panel of 102 AIMS.
SGR: Plot of segregation ancestry in the sample representing expected proportions of ancestry in each individual.
MLE: Estimates of genomic ancestry obtained with MLE.
NP-I-B: Estimates of ancestry in STRUCTURE using no ancestral populations and independent alphas.
NP-S-B: Estimates of ancestry in STRUCTURE using no ancestral populations and same alphas.
Figure 5.1 Continued
F-I-B: Estimates of ancestry in STRUCTURE using ancestral populations and specifying the ‘Popflag’ option and independent alphas option.
F-S-B: Estimates of ancestry in STRUCTURE using ancestral populations, specifying the ‘Popflag’ option and same alphas option.
NF-I-B: Estimates of ancestry in STRUCTURE using ancestral populations, not specifying the ‘Popflag’ option but including independent alphas option.
NF-S-B: Estimates of ancestry in STRUCTURE using ancestral populations, not specifying the ‘Popflag’ option and specifying the same alphas option.
Figure 5.2: Triangle plots of ancestry estimates in simulated samples using the unbalanced panel of 100 AIMS.

SGR-UB: Plot of segregation ancestry in the sample representing expected proportions of ancestry in each individual.
MLE-UB: Estimates of genomic ancestry obtained with MLE.
NP-I-UB: Estimates of ancestry in STRUCTURE using no ancestral populations and independent alphas.
NP-S-UB: Estimates of ancestry in STRUCTURE using no ancestral populations and same alphas.
F-I-B: Estimates of ancestry in STRUCTURE using ancestral populations and specifying the ‘Popflag’ option and independent alphas option.
F-S-B: Estimates of ancestry in STRUCTURE using ancestral populations, specifying the ‘Popflag’ option and same alphas option.
NF-I-B: Estimates of ancestry in STRUCTURE using ancestral populations, not specifying the ‘Popflag’ option but including independent alphas option.
NF-S-B: Estimates of ancestry in STRUCTURE using ancestral populations, not specifying the ‘Popflag’ option and specifying the same alphas option.

**Effect of using a balanced vs. unbalanced marker panel**

The most striking effect appears to be that due to the characteristics of the marker panel. As seen from the estimates of MSE in Tables 5.4 and 5.5, all options performed better when a balanced panel of AIMs was used. Correlations between expected and
observed estimates were higher with the balanced panel compared to those observed using the unbalanced panel for all options. The results of the KS test indicate that ML estimates with a balanced panel are most consistent with expected estimates, while in the unbalanced panel there are significant departures on both the EU and IA axes. Cumulative percentile plots further indicate the departure of ancestry estimates from expected values as shown in Figure 5.3. For most individuals, IA estimates are significantly higher than expected and EU estimates are significantly lower than expected using the unbalanced panel, reflecting the low distinguishing power between EU-IA in this panel.
Figure 5.3: Cumulative Percentile plots of distribution of ancestry for each axis in MLE-UB and MLE-B. Solid line represents expected segregation ancestry and broken line represent estimated genomic ancestry for West African (WAF), European (EU) and Indigenous American ancestry (IA). MLE-UB: ML estimates with unbalanced marker panel; MLE-B: ML estimates with balanced marker panel.
Using STRUCTURE, the same general trend is observed. Across all options, using the balanced panel of markers provides estimates which deviate least from expected estimates. However, IA ancestry is underestimated in all options except when the STRUCTURE model has no ancestral populations and uses the same alpha parameter (models NP-S-B and in NP-S-UB), in which case IA ancestry is overestimated.

**Effect of using ancestral populations in STRUCTURE**

Although STRUCTURE can make inferences in the absence of ancestral populations, the utility of including some ancestral individuals is immediately apparent from figures 5.1 and 5.2. The options NP-I-B, NP-S-B, NP-I-UB and NP-S-UB performed worst in the categories of balanced and unbalanced markers. Both t-tests and KS tests indicate that means and overall distributions of ancestry estimates differ significantly. This is illustrated in Figure 5.4 which shows estimates for separate alpha options and in Figure 5.5 which shows estimates for same alpha option.
Figure 5.4: Cumulative Percentile plots of distribution of ancestry for each axis in NP-S-UB and NP-S-B. Solid line represents expected segregation ancestry and broken line represent estimated genomic ancestry for West African (WAF), European (EU) and Indigenous American ancestry (IA). For both plots STRUCTURE models included no ancestral populations and same alpha options; NP-S UB: Results with unbalanced panel; NP-S B: Results with balanced panel
Figure 5.5: Cumulative Percentile plots of distribution of ancestry for each axis in NP-I-UB and NP-I-B. Solid line represents expected segregation ancestry and broken line represent estimated genomic ancestry for West African (WAF), European (EU) and Indigenous American ancestry (IA). For both plots STRUCTURE models included no ancestral populations and independent alpha options; NP-I UB: Results with unbalanced panel.
**Effect of using Popflag option in STRUCTURE**

This is an interesting option to consider when using STRUCTURE. When ancestral populations are included in the sample and each population is identified as being a separate population using the ‘Usepopinfo’ option, ‘Popflag’ is a second option that can be used to indicate specifically which of the individuals are ancestral individuals and which are admixed persons, by specifying a Boolean variable. The Boolean variable takes a value of either 0 or 1. All ancestral populations are identified using ‘1’ and all admixed individuals or test subjects are identified using the value ‘0’. This specifically uses the ancestral individuals to set priors for allele frequencies as well as priors for the vector representing admixture proportion (Pritchard et al. 2000).

No significant effects of specifying the Popflag option is seen when using a balanced marker panel, as all options F-I-B, F-S-B, NF-I-B, NF-S-B have similar values of MSE and correlation coefficients as seen in Table 5.4. However, when the marker panel is unbalanced, the effect of specifying the Popflag option is quite distinct as seen from MSE and correlation values in Table 5.5. The plots for options NF-I-UB and NF-I-B are shown in Figure 5.6 and NF-S-UB and NF-S-B are shown in Figure Figure 5.7 respectively.
Figure 5.6: Cumulative Percentile plots of distribution of ancestry for each axis in NF-I-UB and NF-I-B. Solid line represents expected segregation ancestry and broken lined represent estimated genomic ancestry for West African (WAF), European (EU) and Indigenous American ancestry (IA). STRUCTURE models included ancestral populations but Popflag option was not specified and independent alpha option was used.
Figure 5.7: Cumulative percentile plots of distribution of ancestry for each axis in NF-S-UB and NF-S-B: No Popflag, same alpha and unbalanced marker panel and NF-S-B: No Popflag, same alpha and balanced marker panel. Solid line represents expected plot and dotted line represents observed genomic ancestry.
Effect of using same alpha parameter vs. independent alpha parameter in STRUCTURE

The alpha parameter specifies how admixture is modeled in the different populations. Using separate alphas for each population allows for an asymmetric distribution of Q, the vector for admixture proportions, indicating unequal contributions from different populations. Restricting alphas to be the same for all populations is analogous to modeling equivalent admixture proportions from all populations, which based on our simulation scheme, is inappropriate. While in the balanced marker panel the influence of the alpha parameter is of less consequence, in the unbalanced marker panel, the effects of the alpha parameter are extremely significant, as seen from Table 5.5, Figure 5.2, Figure 5.7 (plot for option NF-S-B) and in Figure 5.8 which shows the option F-S-UB. The confounding result in this category is from the option NF-I-UB, shown in Figure 5.6 (plot: NF-I-UB) and Figure 5.2. Compared to the ‘Same alpha’ model with the remaining options kept the same (i.e. plot of NF-S-UB in Figure 5.2 and in Figure 5.7, this option appears to be performing worse. However, note that the option NF-S-UB performs worse than option NF-S-B, highlighting the importance of the marker panel.
MLE compared to Bayesian methods

Using a balanced panel of markers shows that estimates with MLE deviate least from the expected estimates. T-tests and KS tests on each ancestry axis indicate non-significant differences between expected and observed estimates for the balanced panel (See tables 5.4 and 5.5). However, with an unbalanced panel there are significant differences between expected and observed estimates for both the EU and IA axes. In contrast, using STRUCTURE with F-I-B option, deviations are only observed for the IA axis which is significantly underestimated.
**Discussion**

This study demonstrates some practical considerations that need to be made when estimating individual ancestry proportions using either ML or STRUCTURE. I have used a simple simulation scheme in which it is possible to specify admixture proportions and exactly track the estimate of alleles in each individual that are inherited from each ancestral population over a number of generations. This represents a controlled situation that is unlikely to be reflected in reality, since there will always be uncertainty associated with the data (in terms of ancestral population choice, allele frequency estimates, admixture model). However, using this situation allowed me to make some basic comparisons and investigate the practical effects of several of the parameters. The program STRUCTURE is a popular program used in population genetics to investigate stratification in populations. It has recently been used to expressly classify populations, some with known history of admixture (Tang et al. 2005a), into distinct non overlapping groups, in admixture mapping (Zhu et al. 2005). Other uses have included identifying population groups (DeLisi et al. 2002; Rosenberg et al. 2002; Bamshad et al. 2003). However, as seen from the analyses in the present study, choice of parameters and marker panels used will have a distinct effect on the results obtained. In addition, several authors have shown that ancestry proportions estimated with ML and STRUCTURE are generally consistent with both methods (Gower et al. 2003; Shriver et al. 2003; Bonilla et al. 2004; Parra et al. 2004; Bonilla et al. 2004a; Reiner et al. 2005; Salari et al. 2005; Tang et al. 2005a), but have not commented on the differences in estimates observed. In one study, estimates obtained assuming a two-way admixture model in STRUCTURE have been correlated with estimates obtained with three-way admixture model in ML, and it was shown that the estimates for the major ancestry axis, in this case WAF in African Americans, is highly correlated (correlation coefficient 0.98, P<0.001) (Reiner et al. 2005). In a second study, several models of admixture were used (two-way vs. three-way admixture) following the fact the STRUCTURE analyses indicated presence of two subpopulations in the population sample consisting of Puerto Rican subjects from New York (Bonilla et al. 2004a), although, estimates of ancestry obtained with three-way
admixture model were used in subsequent analyses to test for effects of ancestry on phenotypes. A third study has compared the estimates of individual ancestry obtained with STRUCTURE to those obtained with an iterative ML method (Tang et al. 2005b), to show that estimates obtained with the proposed iterative ML method agrees closely with those obtained with STRUCTURE. However, this report indicates that for running STRUCTURE, all default parameters (which may or may not be appropriate) were used except for length of initial burn in and final runs for parameter estimation. The default options in STRUCTURE are to use no prior population information and assume a uniform prior for the alpha parameter and same alpha parameter for each population, in addition allele frequencies are assumed to be correlated and an admixture model is used. However, the report does not discuss the implications of choosing this particular model. As seen from the present analysis, the choice of the alpha parameter affects the final estimates of ancestry extensively and requires careful consideration.

With the availability of larger marker panels and lower genotyping costs, issues relating to marker choice are gradually becoming less important. However, methodological issues will persist until they are addressed systematically. What is evident from previous chapters in this thesis is that the combined effects of the marker panel and analytical method together are likely to cause some confounding. For instance, using an unbalanced marker panel, not using ancestral populations and specifying same alpha parameter is not the best option to use. Where both of these affect the final estimates, reliability of such estimates is questionable.

This preliminary analysis highlights the importance of using a balanced marker panel, when three ancestral populations have contributed to an admixed sample. In the three previous chapters I have analyzed African Americans assuming three way admixture between WAF, EU and IA populations, which is consistent with the demographic history of the US where all three populations have cohabited for the last 500 years. Although I have observed strong correlation for proportional WAF and EU ancestry, IA ancestry has been estimated to be significantly different by different methods. There is a possibility that the statistical discrepancy has no biological effect, but further research is required to better understand this. Using a balanced marker panel both
ML and the F-I-B option in STRUCTURE perform well, as expected. However, the mean IA ancestry estimated with F-I-B is lower and KS tests indicate small but significant difference from expected proportions (as seen in Figure 5.1). With the unbalanced panel ML overestimated and F-I-B underestimated proportional IA ancestry. Since these are preliminary studies, further research is required to investigate if this pattern of misestimating is consistent, by repeating similar models.

This analysis also highlights the importance of including some ancestral individuals in STRUCTURE which can be used for ascertaining suitable priors, particularly when the marker panel being used is unbalanced. In contrast, with a balanced panel of markers neither the choice of alpha nor the inclusion of the ‘Popflag’ option appears to have any significant effect, which is interesting and requires further investigation. The IA axis was underestimated in all options however. This is important for another reason as well which perhaps has more fundamental implications.

In all of the STRUCTURE analyses extreme observations of ancestry proportions were not observed. The simulated samples, both for the balanced and unbalanced panels have 47 individuals each who have 0.0 admixture from either IA or EU. Using the balanced panel of markers in ML, 37 of these individuals show 1.0 WAF and with the unbalanced panel 39 show 1.0 WAF. In instances where non-WAF ancestry was estimated it was <0.02 in each case. In contrast, none of these Individuals show 1.0 WAF in any of the STRUCTURE models. In the two most appropriate models, F-I-B and F-I-UB, the mean estimates for this group of individuals was 0.97 ± 0.01 for the unbalanced panel and 0.95 ± 0.01 for the balanced panel. These highlight further the importance of which individuals are included in the sample for the STRUCTURE analyses. The simulated samples have 0.25 EU ancestry on average, with some individuals showing higher EU and some showing higher IA. In ML, each individual is analyzed as an independent observation and other individuals included in the sample do not have any influence. However, in STRUCTURE all individuals included in the analysis are used to make inferences for each person. This partly explains why IA is consistently underestimated in the simulated samples. Overall IA ancestry in the samples is low. In both simulated sets ~70 individuals have 0.0 IA ancestry, while most of the remaining
individuals have <0.1 IA ancestry. This appears to be sufficient for the program to assign non-zero IA ancestry to all individuals.

While ML explores the entire parameter space to find the highest likelihood peak, MCMC explores the surface rather than just searching for the highest peak. The iterations after burn-in basically represent different points on the probability distribution. Each iteration of MCMC can identify a certain point on the posterior probability surface that best represents the data, with probability proportional to the relative probability of each point on the surface. Hence the points after the initial burn in represent approximately random samples from the posterior distribution. Taking the mean value across these estimates is a legitimate statistical procedure since the expected value of a parameter is the mean. Theoretically even in the MCMC approach to estimation, the maximum posterior probability point is still the best and given the assumed prior provides most information that the data set can give about the underlying truth. Whether using the mean or the maximum as an estimator has any biological significance requires further analysis.

Through this simulation study I have explored some of the issues that affect admixture estimates. This is a preliminary step and further extensive analyses are required to understand overall effects of markers and methods on estimates. One negative aspect in this study could be that I have simulated a perfect situation where there is no ambiguity in either the choice of ancestral populations or estimates of ancestral allele frequencies. I have also not investigated the effects of using a smaller marker panel, as has been done for most practical studies to date (for example using a panel of 40 AIMs). This situation is appropriate as a first pass to investigate the behavior of the programs under the perfect situation, although obtaining such a situation in reality may not be possible. The obvious next step is to use the basic simulation scheme and vary different parameters. For instance, rarely will it be possible to obtain samples of specific ancestral populations that have contributed to an admixed population. Drift is likely to have influenced both the ancestral populations and the admixed population to varying extents. Since Bayesian methods are designed to effectively analyze such situations, it is expected that estimates obtained with Bayesian methods will be more reliable.
References


structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 76:268-275
Chapter 6

Conclusions and Future Directions

‘I note the obvious differences
in the human family

Some of us are serious,
some thrive on comedy.

… In minor ways we differ,
in major we’re the same….

I note the obvious differences
between each sort and type,
but we are more alike, my friends
than we are unalike.”

----- Maya Angelou (from The Human Family)

Introduction

Understanding genetic variation within and between populations is important and is one of the most challenging and often controversial aspects in human population genetic research. It is well documented that several phenotypes vary between population groups. Several studies have documented the racial/ethnic differences between populations suggesting a genetic basis underlying these phenotypes. Investigating these differences is likely to provide important information on the mechanisms that underlie the biology of the trait. Populations with a history of admixture offer a unique opportunity in that they represent a ‘natural experiment’ in which it may be possible to test the ancestry
of alleles and investigate ancestry-phenotype correlations in order to establish a genetic basis for diseases and traits. Recently, a lot of attention has been focused on investigating whether “Race” has a biological basis or not and the appropriateness of applying race as a discrete variable to define genetic differences between populations. It is important to recognize that Race or ethnicity has a broader cultural connotation. It is possible, that some genetic differences do exist between population groups, reflecting the effects of genetic drift and natural selection since the time populations separated. However, the extent of such differences is unknown. Understanding these differences and being able to measure them in a meaningful way is necessary. Recently admixed populations are an important resource for studying the phenotypic effects of genetic differences between populations. However, great caution needs to be taken to prevent misrepresentation and over simplification of the results. Several recent studies have reported genetic differences between racially/ethnically defined population groups, that appear to indicate that such labels represent distinct non-overlapping genetic clusters that should be identified, but once such clusters have been identified, variations within groups are of lesser consequence (Rosenberg et al. 2002; Jorgenson et al. 2005; Tang et al. 2005). While studying differences is important and useful, it has to be recognized that identifying discrete groups based on genetic similarity alone and correlating these differences with racial/ethnic group labels could be misleading. As this thesis demonstrates, substantial genetic variations exist within groups and these groups overlap in ways to create a continuous distribution of genomic ancestry proportions. Understanding what genomic ancestry is and how it can be used to answer some biological questions is important, especially in today’s world, where population differences (assumed or real: behavioral, cultural, social and may be even biological) appear to play such an important role in our daily lives.

In this thesis I have focused on understanding how genomic ancestry varies among and between culturally identified population groups. I have investigated different methods that are currently available for measuring genomic ancestry. To establish a genetic basis for phenotypes that vary between population groups I have examined
ancestry component-phenotype correlations. Detecting such correlations is an intermediate step towards mapping genes that underlie these phenotypes. Using a small sample, I have performed admixture mapping of some complex phenotypes. In the process I have encountered several important issues that I discuss below. In addition, I discuss prospects for future research.

Ancestry estimation and variation of ancestry proportions in different populations

Estimates of individual ancestry are useful measures of genetic variability and should be incorporated in genetic association studies to control for false positives and false negatives arising due to admixture stratification. The aim of the research reported in this thesis is to estimate individual ancestry proportions in different populations and observe how individual ancestry varies across populations. From these studies it appears that using different groups or clusters is not the best way of representing the genetic variation within and between populations, but rather as a continuum of individual admixture proportions. Thus controlling for race or genetic cluster would still put studies at risk of false positives or false negatives unless interindividual variation within and between groups is accounted for. As seen in Chapters 2, 3 and 4 individual estimates of ancestry show a wide range of variation, especially in populations with known histories of admixture like African Americans and Mexican Americans. Even populations which are otherwise assumed to have relatively less admixture, like the different European-American populations used in these studies, have some variation that may be attributed to individual genomic ancestry. It should be pointed out that I have focused on the genomic ancestry that is related to populations which have genetic roots indifferent continents. I have not examined variation within continents and this could be an important aspect to consider. For instance, to better understand genetic variation within European Americans it will be important to understand genetic variation within the European continent. Overall, understanding genetic variation is important, especially in the quest for genes
that underlie complex phenotypes. It is well documented that several complex phenotypes and diseases show population specific variation. Investigating these differences may provide an understanding of the biology underlying these phenotypes. Several factors need to be considered when computing estimates of genomic ancestry that are discussed below.

**Considerations for individual genomic ancestry estimation:**

Several issues need to be considered for computing reliable estimates of genomic ancestry. The choice of marker panels, methods, ancestral populations, sample sizes and admixture model all have an impact on the final estimate obtained, as illustrated in this thesis. Of the conditions listed above, choice of markers is the most critical issue. Given the recent origin of our species most loci are not informative for ancestry. Loci with high allele frequency differences between populations are appropriate for ancestry estimation and have been designated as ancestry informative markers (AIMs). In the past few years several AIM panels have been developed that should be used for individual ancestry estimation (Zimmet et al. 1977; Smith et al. 2001; Akey et al. 2002; Shriver et al. 2003; Shriver et al. 2004; Smith et al. 2004; Shriver et al. 2005). Several measures of informativeness have been proposed, as described in Chapter 1. Apart from allele frequency difference (δ) or its multiallelic extension, composite delta (δ_c), ancestry information can be measured by score variance ‘f’ (Fisher information) (McKeigue 1998; McKeigue 2000; Molokhia et al. 2003; Shriver et al. 2003; Pfaff et al. 2004) or by the expected log likelihood ratio or Kullback-Leibler information (Rosenberg et al. 2003; Smith et al. 2004) which is contributed by typing a gamete at the marker locus. All these measures can be calculated from ancestry-specific allele frequencies and are expressed as the proportion of information that would be extracted by a perfectly informative marker, against a uniform prior distribution. The symbols “f” (McKeigue 1998) and I_n (Rosenberg et al. 2003) have been used to designate Fisher information and Kullback-Leibler
information, respectively. Both methods rank markers similarly for ancestry information content although the absolute values for such measures differ (McKeigue 2005, McKeigue personal communication). Fisher information content “f” is related to locus-specific $F_{ST}$ distance and is a value representative of the differentiation between two populations at a single locus and is equivalent to Wahlund’s standardized variance for allele frequency. Simulation studies for estimating the information content of markers with varying levels of $f$ have shown that for efficiently estimating individual ancestry with a standard error of no more than 0.1 we would require at least 40 unlinked AIMs with average $\delta$ of 0.67 (McKeigue et al. 2000). Microsatellites are less useful in this case since assigning ancestry to rare alleles or common alleles can be confounding. For instance where alleles are present in one population but not in another, or when some alleles have similar frequencies in different populations, inference of allelic ancestry will be difficult. Although composite $\delta$ ($\delta_c$) has been used in chapters 2 and 4 as a measure of ancestry informativeness, this is an aggregate measure over all alleles at a locus with little information for any particular allele. Especially, with a smaller panel, the uncertainty associated with such estimates can be high. This is evident when comparing the estimates of Bayesian methods to those obtained with maximum likelihood. The discrepancy between estimates obtained with the two methods was greatest for the results in Chapter 2 (where a panel of 37 microsatellites and 1 SNP AIM were used for inferring individual ancestry), compared with those in Chapter 3 (Panel of 34 SNP AIMs), Chapter 4 (Panel of 311 microsatellites), Chapter 5 (Simulated data with a panels of 100 AIMs and 102 AIMs) or Appendix D (11,555 SNPs from the Affymetrix 10K chip).

A more critical and less frequently discussed issue is perhaps the choice of using a balanced marker panel, one in which there are equal numbers of markers that provide information for alternate axes of ancestry. This issue is likely to become more important as additional populations are analyzed, where more than two populations have contributed. For example, in some Latino populations, and in populations with more complex history of admixture, like the Trinidadians where there is evidence of admixture between five ancestral populations, Europeans, West Africans, Indigenous Americans, Chinese and Indians (Molokhia and McKeigue 2000).
Choice of ancestral populations and sample sizes of these populations will also be important. Several European, West African and Indigenous American populations have contributed to contemporary European American, African American and Latino populations. To encompass most of the variation in ancestral populations we will require to sample from multiple populations. The data used in Chapter 3 is most appropriate in terms of ancestral populations and sample sizes. Ancestral allele frequencies were estimated from samples of 242 Indigenous Americans (Maya, Southwestern US Indigenous Americans: Cheyenne, Pima and Pueblo), 241 Europeans (British, Irish, German and Spanish), 545 West Africans (from Sierra Leone, Nigeria, Central African Republic, Congo, Zambia and Uganda) for all markers. However, such extensive panels of ancestral allele frequencies have only been established for few markers (Parra et al. 1998; Shriver et al. 2003; Bonilla et al. 2004b; Smith et al. 2004; Bonilla et al. 2004a). It is also notable that with allele frequencies for which there is adequate information, both Bayesian and classical methods show remarkable concordance, both in proportional ancestry estimation and in admixture mapping [see for example, the results of Chapter 3, (Shriver et al. 2003), and simulations with balanced panel of markers in Chapter 5]. As opposed to this, for Chapter 2, ancestral allele frequencies were estimated from one group each from the three ancestral populations, in Chapter 5, two ancestral European, two ancestral Indigenous American and three populations samples from different parts of West Africa were used. Even with the 10K panel used in the analysis shown in Appendix D (Shriver et al. 2005), only one ancestral population has been used to represent European (Spanish) and West African (Mende) populations respectively and two (Nahua and Quechua) for Indigenous Americans. This choice could partly explain the movement of European American individuals away from the European vertex in the analysis of the 10K data set (Appendix D, Figure 1 and Figure 3). Although, these analyses provide much needed empirical evidence for variation in proportional ancestry, the potential drawbacks should be considered for each case when interpreting the results. The choice of ancestral populations used will also have a large impact on estimates of ancestry. It is likely (maybe obvious in some instances) that genetic drift has affected both the admixed populations and descendants of ancestral populations which contributed to the admixed
population in the time since the admixture event/s. This is going to be of greater consequence when using the ML method since in this method ancestral allele frequencies are treated as fixed, invariant constants based on which inferences are derived. In contrast, theoretically, Bayesian methods can handle misspecifications more appropriately, since inferences are based on both ancestral and admixed samples. In the program ADMIXMAP (McKeigue et al. 2000; Hoggart et al. 2003; Hoggart et al. 2004; McKeigue 2005), specific tests have been designed to investigate the effects of using allele frequencies which have been specified incorrectly. Further analysis, through simulations is required to evaluate the efficacy and usefulness of such tests. Analyses are currently in progress to examine the effects of incorrectly specifying ancestral allele frequencies and will provide important information regarding the reliability of estimates obtained with each method. While sample sizes of ancestral populations are important for maximum likelihood and Bayesian methods, the size of the admixed population sample being investigated will also impact the estimates obtained from Bayesian methods which make inferences using the admixed sample as well.

The issues surrounding choice of methods for individual ancestry estimation have been dealt with extensively in Chapter 5. With lots of AIMs estimates obtained with either the classical or the Bayesian methods the estimates are expected to be asymptotically equal. In addition, it is important to understand the different modeling options when using the Bayesian inferences applied in the STRUCTURE and ADMIXMAP programs. Preliminary analyses are presented in Chapter 5, where effects of marker panels and some of the modeling options have been investigated. Studies are currently underway to investigate several other modeling options in STRUCTURE and ADMIXMAP programs (for instance, using correlated vs. uncorrelated allele frequencies in STRUCTURE; using assortative vs. random mating models in ADMIXMAP). It is remarkable that using a balanced marker panel appears to negate the effects of several of the modeling options in STRUCTURE, as long as some ancestral populations are included in the analysis along with the admixed individuals. The simulation studies also show that using an unbalanced panel in conjunction with some of the modeling assumptions can cause substantial adverse effects, for example using a model assuming
same alpha parameter for all individuals, not using the ‘Popflag’ option to designate ancestral individuals and using an unbalanced panel caused proportional ancestry estimates to deviate significantly from expected values. Better understanding is required of the effects of each parameter prior to using the estimates of ancestry in further analyses.

The choice of admixture model assumed will have substantial effects in some cases as well. In the absence of documented evidence it becomes difficult to decide \textit{a priori} what population model should be assumed. For instance, different studies have modeled admixture in African Americans as either a two-way admixture model between Europeans and West Africans (Hoggart et al. 2003; Patterson et al. 2004; Reiner et al. 2005; Zhu et al. 2005) or as a three way model assuming admixture between Europeans, West Africans and Indigenous Americans (Shriver et al. 2003) and Chapters 2, 3 and 4 of this thesis). It is not possible to say \textit{a priori} which model is more appropriate and some demographic characteristics will have to be used to make reasonable assumptions. It will be necessary to test for appropriateness of an ancestral population being used for each sample using available methods (Bonilla et al. 2004b).

More interesting are the effects of low levels of admixture from a third population in presence of greater admixture from two populations. Further analysis is required to investigate the effects of incorrectly specifying the admixture model and its consequences. For instance, what are the effects of using a two-way admixture model when more than two populations have contributed to the population or what the effects are of specifying a three-way admixture model when only two populations have contributed? This is the first step in a series of such analyses and more complex models involving four or more populations will be examined in future.

\textbf{Ancestry component-phenotype associations}

As illustrated in Chapters 3 and 4, proportional estimates of ancestry can be used to investigate the genetic basis of disease and various phenotypes. When phenotypes are
affected by multiple factors including genes, environment and behavior, distinguishing the effects of different components is difficult, since effects of one component may be influenced and related to effects of a second component. When population differences in phenotypes can be attributed to non genetic and genetic causes, estimates of proportional ancestry will be useful to investigate the genetic basis of the trait. If proportional ancestry varies with trait/disease phenotype it may be possible to establish a genetic basis for the trait or diseases. Strong ancestry-phenotype correlations have been reported for skin pigmentation in two populations of African descent ($R^2 = 0.21 \ P<0.0001$ for African American sample and $R^2 = 0.16, P<0.0001$ in African Caribbean sample) (Shriver et al. 2003), one Latino population from Colorado ($R^2 = 0.082, P<0.001$) (Bonilla et al. 2004b) and one small sample of Puerto Rican women from New York City ($R^2 = 0.59, P<0.001$) (Bonilla et al. 2004a). Preliminary ancestry correlations have also been reported for BMI, plasma glucose and prevalence of type 2 diabetes in Pima Indians (Williams et al. 2000), bone mineral density in Puerto Ricans (Bonilla et al. 2004a), BMI in an African American sample (Fernandez et al. 2003), Insulin sensitivity acute insulin response to glucose in an African American population (Gower et al. 2003) and asthma in a Mexican American population ($R^2 = 0.21, P = 0.005$) (Salari et al. 2005). However, the $R^2$ values associated with complex traits are typically much lower. It is unlikely that very high $R^2$ values will be observed for complex phenotypes, especially in the absence of appropriate environmental or behavioral covariates. This is seen from the results of chapters 3 and 4. Although the correlations observed are statistically significant, the strength of these associations is weak in some cases, as measured by $R^2$. But observing such associations are important since they may help decompose different factors that affect the phenotypes in some populations. Consider that ancestry-phenotype correlations were only observed in populations with known history of substantial admixture, but not in populations where overall admixture is low. In Chapter 3 systolic blood pressure shows a suggestive negative correlation with proportional European American ancestry (EU) in a sample of African American individuals ($R^2 = 0.28$ and $P = 0.047$), while no correlations were observed between proportional West African ancestry (WAF) or proportional Indigenous American ancestry (IA). In Chapter 4, BMI and blood glucose show correlations with
ancestry in Mexican Americans ($R^2 = 0.11, P < 0.0001$ for both EU and IA ancestry for BMI, and $R^2 = 0.03, P = 0.006$ for fasting blood glucose for EU and IA ancestry). In the combined African-American samples described in Chapter 4, BMI shows a suggestive negative association with WAF ancestry ($R^2 = 0.03$ and $P = 0.055$) and separately fasting blood glucose shows a suggestive association with IA ancestry ($R^2 = 0.1, P = 0.04$). Interestingly, when only females were analyzed, a suggestive association between glucose and EU ancestry was observed in the African-American female subjects from the GENOA study ($R^2 = 0.07, P = 0.05$). It is also noted that while in Mexican Americans estimates obtained with both maximum likelihood and STRUCTURE have shown an association with the phenotypes, in African Americans only the estimates obtained with maximum likelihood show an association. In some previous reports where both maximum likelihood and STRUCTURE have been used to infer proportional ancestry, ancestry-phenotype associations have only been examined using the results of maximum likelihood analysis (Shriver et al. 2003; Bonilla et al. 2004b; Bonilla et al. 2004a; Reiner et al. 2005). The results from Chapter 4 also highlight the issues related to model choice and method used for inferring ancestry. Typically, using STRUCTURE the estimates of ancestry from the axis with the least contribution are underestimated. This probably explains the lower estimates of IA ancestry in African Americans and reduced estimates of WAF ancestry in Mexican Americans that is observed with STRUCTURE, compared to the estimates obtained with ML. While the statistical differences are intriguing, it is not immediately apparent whether these discrepancies will influence our understanding of the biological effects when we use estimates obtained with one method to test for ancestry component-phenotype associations. If estimates obtained with either method are correlated and both are reliable then it should be possible to use either estimate to infer an ancestry-phenotype association. This issue highlights the importance for understanding what, if any, the implications might be, for choosing estimates obtained with one method over another. When some associations are only observed with one method and not with another, it could indicate and possibly dictate which method should be used for inferring proportional ancestry when these estimates of ancestry are to be used for investigating further ancestry-phenotype associations. Additional simulation studies are required to
further explore these effects and for making suitable recommendations. The simulations presented in Chapter 5 represent an appropriate situation in terms of ancestral population choice, sample size and marker panel used. For the unbalanced panel, markers were only selected based on WAF-EU difference with no consideration given to IA informativeness. While for the balanced panel information for each ancestral population was specifically included, to prevent any skewing due to the marker information content. Of special significance is the issue of incorrectly specifying ancestral allele frequencies, as has been discussed previously.

When investigating any complex phenotype-ancestry correlations, it should be remembered that a straightforward, linear relationship may not exist, and considering the effects of environmental and behavioral aspects is of vital importance. It is unlikely that relationships as strong as those observed for skin pigmentation will be observed, since few traits are as different among populations and effects of non-genetic factors may be considerable. However, where present these correlations will help to establish the genetic basis of the phenotypes. With the availability of large, well characterized AIM panels choosing markers for an analysis are easier. Therefore, new analyses should be considered by specifically using appropriate panels and including suitable non-genetic covariates.

Admixture mapping

In the quest for genes underlying complex phenotypes admixture mapping is a relatively new concept that has recently garnered a lot of attention (Halder and Shriver 2003; McKeigue 2005; Smith and O'Brien S 2005), since it was first proposed that the long range allelic associations observed in admixed populations could be used for mapping disease genes (Chakraborty and Weiss 1988). For complex diseases that are not amenable to family based linkage analyses (due to reasons including allelic and locus heterogeneity, phenotypic heterogeneity, late age of onset, incomplete penetrance),
association studies have been proposed as an alternative method. However, population stratification is one issue that needs careful consideration, since undetected and uncontrolled stratification will lead to false positives and false negatives. While controlling for stratification is important, using the interindividual variation in ancestry proportions will be necessary for admixture mapping. It should be clarified, that the method originally proposed by Chakraborty and Weiss (1986) is what became known as Mapping by Admixture Linkage Disequilibrium (MALD) (Stephens et al. 1994). A second approach that attempts to assign ancestry to alleles at each locus and is more analogous to linkage analysis in an experimental cross is referred to as Admixture Mapping (AM). Although both methods aim to use the allelic associations, statistical treatment and the underlying conceptual framework of how the process of admixture is used by these methods is slightly different. Two recent reports have each attempted to explore these two situations (McKeigue 2005; Smith and O'Brien S 2005). McKeigue describes the underlying theory of admixture mapping. This is the general method used in the programs ADMIXMAP (McKeigue et al. 2000; Hoggart et al. 2003; Hoggart et al. 2004) and also in the programs ANCESTRYMAP (Patterson et al. 2004) and MALDOSOFT (Montana and Pritchard 2004), although the specific statistical treatments are different. Smith and O’Brien have described the approach that uses admixture linkage disequilibrium (ALD). However, none of the available methods actually focus on ALD, instead the focus is now on ascertaining ancestry of alleles at each locus. Further analysis and investigation is required to better understand these methods. It should be cautioned that all methods are statistically and computationally extremely sophisticated and in some cases are still being refined and extended. Therefore, care and caution will need to be applied when using any of these programs and special attention needs to be paid in order to avoid over interpretation of results.

In this context it is important to point out a recent study on admixture mapping of hypertension (Zhu et al. 2005). The markers used in this study are the same as those that were used in the analyses described in Chapter 4. Thus, the overall comments about insufficient power to infer informativeness of alleles at a locus are applicable to this study. The study by Zhu et al., (2005) relies on estimating locus specific ancestry
separately in cases and controls, using STRUCTURE and uses these estimates in a
second analysis to test for significant differences in locus ancestry between cases and
controls. In addition, the authors have assumed that admixture in African Americans is
adequately described by a two-way model of admixture between Europeans and West
Africans. Finally, except for years of education no other information was used to include
non-genetic factors. For a phenotype as complex as hypertension, it is unlikely that
environmental and behavioral factors do not affect the phenotype, yet these issues were
not addressed adequately. Samples across different geographical regions were used in the
same analysis, as has also been done in Chapters 3 and 4 of this thesis. While individual
estimates of ancestry may account for the genetic heterogeneity, combining samples
across regions and not taking into account associated environmental variables
considerably undermines the study. This, in conjunction with an inadequate marker panel
is likely to have affected the overall analysis.

What then constitutes an effective admixture mapping study? An important factor
will be to use a large, well characterized panel of AIMs and several such panels are now
available (Smith et al. 2001; Shriver et al. 2004; Shriver et al. 2005) and several others
are being currently developed. It is unlikely that one panel will be sufficient for all
admixture mapping studies since the dynamics of admixture will differ from population
to population. The markers are selected based on expected theoretical measures and only
through practical applications of the markers will their effectiveness be truly evaluated. A
two stage mapping study, one when the primary screening is achieved with ~ 1000
genome spanning panel followed by a second, denser panel typed in regions of interest
(Hoggart et al. 2004; McKeigue 2005) is a reasonable plan to follow. In addition to
identifying new markers, more ancestral populations need to be ascertained and allele
frequencies from multiple ancestral populations need to be assembled. Another important
issue will be to better characterize the phenotypes of interest. For instance, using
intermediate phenotypes (like systolic or diastolic blood pressure), as opposed to using
presence or absence of a disease (like hypertension vs. normal blood pressure) may be an
alternate approach to consider, focusing on components of the complex outcome to better
understand the biology of the trait. Lastly, a concerted effort needs to be made towards
understanding the non-genetic influences on the trait and use them in genetic association studies. For instance, in Chapter 3 I have combined two samples from different locations, Bogalusa and Philadelphia based on similarities in proportional ancestry levels. Years of education may provide some information regarding variability in socioeconomic status, but there could be several other environmental and behavioral factors that probably have affected the outcomes. Considering several factors together is statistically challenging, but is probably the best way to understand the underlying biology.

In conclusion, in this thesis I have investigated how genomic ancestry can be measured and used to investigate genetic differences between populations. It is expected that these differences are probably subtle for most genetic systems although there may be some with dramatic effects. As stated in the beginning of the chapter, we as a species ‘are more alike, than we are unalike’ and even when searching for differences between populations, this is something that needs to be remembered. As shown from the three studies, the genetic variations observed follow gradients and are not discrete and non-overlapping. In conjunction with the genetic differences we need to consider other cultural, social and behavioral aspects as well, if we want to understand the biology of complex traits.
References


Appendix A

Microsatellite alleles frequencies used for analysis in Chapter 2

<table>
<thead>
<tr>
<th></th>
<th>EU</th>
<th>WAF</th>
<th>IA</th>
<th>AS</th>
<th>EACT</th>
<th>AACT</th>
<th>LCT</th>
<th>LCA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CSF1PO</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>309</td>
<td>0</td>
<td>0.067</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.043</td>
<td>0.02</td>
<td>0.006</td>
</tr>
<tr>
<td>313</td>
<td>0</td>
<td>0.093</td>
<td>0</td>
<td>0</td>
<td>0.004</td>
<td>0.068</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>317</td>
<td>0</td>
<td>0.047</td>
<td>0.051</td>
<td>0.033</td>
<td>0.019</td>
<td>0.034</td>
<td>0.03</td>
<td>0.022</td>
</tr>
<tr>
<td>321</td>
<td>0.273</td>
<td>0.227</td>
<td>0.199</td>
<td>0.237</td>
<td>0.288</td>
<td>0.267</td>
<td>0.21</td>
<td>0.225</td>
</tr>
<tr>
<td>325</td>
<td>0.327</td>
<td>0.193</td>
<td>0.261</td>
<td>0.25</td>
<td>0.294</td>
<td>0.244</td>
<td>0.295</td>
<td>0.298</td>
</tr>
<tr>
<td>329</td>
<td>0.345</td>
<td>0.293</td>
<td>0.415</td>
<td>0.37</td>
<td>0.318</td>
<td>0.287</td>
<td>0.375</td>
<td>0.399</td>
</tr>
<tr>
<td>333</td>
<td>0.055</td>
<td>0.08</td>
<td>0.057</td>
<td>0.1</td>
<td>0.064</td>
<td>0.052</td>
<td>0.04</td>
<td>0.051</td>
</tr>
<tr>
<td>337</td>
<td>0</td>
<td>0</td>
<td>0.017</td>
<td>0.007</td>
<td>0.012</td>
<td>0.005</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>341</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>D2S133</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>304</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>308</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>312</td>
<td>0.036</td>
<td>0.045</td>
<td>0</td>
<td>0.034</td>
<td>0.043</td>
<td>0.059</td>
<td>0.044</td>
<td>0.011</td>
</tr>
<tr>
<td>316</td>
<td>0.295</td>
<td>0.065</td>
<td>0.108</td>
<td>0.074</td>
<td>0.22</td>
<td>0.111</td>
<td>0.235</td>
<td>0.218</td>
</tr>
<tr>
<td>320</td>
<td>0.089</td>
<td>0.045</td>
<td>0.023</td>
<td>0.094</td>
<td>0.081</td>
<td>0.034</td>
<td>0.074</td>
<td>0.069</td>
</tr>
<tr>
<td>324</td>
<td>0.116</td>
<td>0.214</td>
<td>0.273</td>
<td>0.151</td>
<td>0.121</td>
<td>0.173</td>
<td>0.162</td>
<td>0.247</td>
</tr>
<tr>
<td>328</td>
<td>0.179</td>
<td>0.052</td>
<td>0.182</td>
<td>0.094</td>
<td>0.139</td>
<td>0.064</td>
<td>0.137</td>
<td>0.115</td>
</tr>
<tr>
<td>332</td>
<td>0.045</td>
<td>0.136</td>
<td>0.011</td>
<td>0.034</td>
<td>0.034</td>
<td>0.152</td>
<td>0.034</td>
<td>0.04</td>
</tr>
<tr>
<td>336</td>
<td>0.018</td>
<td>0.162</td>
<td>0.085</td>
<td>0.107</td>
<td>0.025</td>
<td>0.116</td>
<td>0.049</td>
<td>0.023</td>
</tr>
<tr>
<td>340</td>
<td>0.054</td>
<td>0.13</td>
<td>0.25</td>
<td>0.201</td>
<td>0.106</td>
<td>0.109</td>
<td>0.093</td>
<td>0.121</td>
</tr>
<tr>
<td>344</td>
<td>0.071</td>
<td>0.052</td>
<td>0.045</td>
<td>0.114</td>
<td>0.113</td>
<td>0.093</td>
<td>0.108</td>
<td>0.075</td>
</tr>
<tr>
<td>348</td>
<td>0.071</td>
<td>0.058</td>
<td>0.023</td>
<td>0.087</td>
<td>0.099</td>
<td>0.064</td>
<td>0.054</td>
<td>0.069</td>
</tr>
<tr>
<td>352</td>
<td>0.027</td>
<td>0.013</td>
<td>0</td>
<td>0.01</td>
<td>0.015</td>
<td>0.016</td>
<td>0.01</td>
<td>0.011</td>
</tr>
<tr>
<td>356</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>360</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gene</td>
<td>59</td>
<td>77</td>
<td>87</td>
<td>153</td>
<td>650</td>
<td>228</td>
<td>101</td>
<td>89</td>
</tr>
<tr>
<td>------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>D3S135</td>
<td>108</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>112</td>
<td>0</td>
<td>0.026</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>116</td>
<td>0</td>
<td>0.006</td>
<td>0.011</td>
<td>0.007</td>
<td>0.006</td>
<td>0.009</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>0.068</td>
<td>0.104</td>
<td>0.069</td>
<td>0.042</td>
<td>0.131</td>
<td>0.112</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>0.271</td>
<td>0.292</td>
<td>0.586</td>
<td>0.327</td>
<td>0.232</td>
<td>0.303</td>
<td>0.312</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>0.271</td>
<td>0.331</td>
<td>0.247</td>
<td>0.297</td>
<td>0.272</td>
<td>0.344</td>
<td>0.252</td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>0.195</td>
<td>0.195</td>
<td>0.08</td>
<td>0.235</td>
<td>0.202</td>
<td>0.178</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>136</td>
<td>0.178</td>
<td>0.039</td>
<td>0.006</td>
<td>0.075</td>
<td>0.141</td>
<td>0.053</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>0.017</td>
<td>0.006</td>
<td>0</td>
<td>0.013</td>
<td>0.012</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>144</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.002</td>
<td>0</td>
<td>0.005</td>
</tr>
<tr>
<td>D5S818</td>
<td>135</td>
<td>0</td>
<td>0.007</td>
<td>0.08</td>
<td>0.036</td>
<td>0.001</td>
<td>0.002</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>139</td>
<td>0</td>
<td>0.035</td>
<td>0</td>
<td>0</td>
<td>0.004</td>
<td>0.059</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>143</td>
<td>0.034</td>
<td>0.014</td>
<td>0.109</td>
<td>0.069</td>
<td>0.045</td>
<td>0.015</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>147</td>
<td>0.069</td>
<td>0.049</td>
<td>0.017</td>
<td>0.217</td>
<td>0.058</td>
<td>0.039</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>151</td>
<td>0.388</td>
<td>0.222</td>
<td>0.563</td>
<td>0.316</td>
<td>0.343</td>
<td>0.219</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>155</td>
<td>0.259</td>
<td>0.326</td>
<td>0.184</td>
<td>0.23</td>
<td>0.359</td>
<td>0.39</td>
<td>0.333</td>
</tr>
<tr>
<td></td>
<td>159</td>
<td>0.241</td>
<td>0.313</td>
<td>0.046</td>
<td>0.115</td>
<td>0.179</td>
<td>0.257</td>
<td>0.127</td>
</tr>
<tr>
<td></td>
<td>163</td>
<td>0.009</td>
<td>0.028</td>
<td>0</td>
<td>0.013</td>
<td>0.01</td>
<td>0.011</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>167</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.002</td>
<td>0.007</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>171</td>
<td>0</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D7S820</td>
<td>255</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>259</td>
<td>0.045</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.016</td>
<td>0.007</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>262</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>263</td>
<td>0.152</td>
<td>0.253</td>
<td>0.063</td>
<td>0.14</td>
<td>0.167</td>
<td>0.211</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>267</td>
<td>0.116</td>
<td>0.071</td>
<td>0</td>
<td>0.083</td>
<td>0.15</td>
<td>0.134</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>268</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>270</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>271</td>
<td>0.313</td>
<td>0.312</td>
<td>0.159</td>
<td>0.167</td>
<td>0.272</td>
<td>0.329</td>
<td>0.312</td>
</tr>
<tr>
<td></td>
<td>275</td>
<td>0.223</td>
<td>0.227</td>
<td>0.352</td>
<td>0.357</td>
<td>0.217</td>
<td>0.211</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td>276</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>279</td>
<td>0.134</td>
<td>0.104</td>
<td>0.33</td>
<td>0.21</td>
<td>0.134</td>
<td>0.095</td>
<td>0.144</td>
</tr>
<tr>
<td></td>
<td>283</td>
<td>0.018</td>
<td>0.013</td>
<td>0.091</td>
<td>0.03</td>
<td>0.039</td>
<td>0.009</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>287</td>
<td>0</td>
<td>0.006</td>
<td>0.006</td>
<td>0</td>
<td>0.004</td>
<td>0.002</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>291</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>
</tr>
<tr>
<td><strong>D8S117</strong></td>
<td>59</td>
<td>74</td>
<td>87</td>
<td>153</td>
<td>648</td>
<td>227</td>
<td></td>
<td></td>
</tr>
<tr>
<td>124</td>
<td>0.025</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.002</td>
<td>0.01</td>
<td>0.011</td>
</tr>
<tr>
<td>128</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.014</td>
<td>0.009</td>
<td>0.015</td>
<td>0.011</td>
</tr>
<tr>
<td>132</td>
<td>0.076</td>
<td>0.02</td>
<td>0.04</td>
<td>0.141</td>
<td>0.098</td>
<td>0.024</td>
<td>0.088</td>
<td>0.062</td>
</tr>
<tr>
<td>136</td>
<td>0.059</td>
<td>0.027</td>
<td>0.069</td>
<td>0.17</td>
<td>0.063</td>
<td>0.042</td>
<td>0.069</td>
<td>0.073</td>
</tr>
<tr>
<td>140</td>
<td>0.119</td>
<td>0.095</td>
<td>0.109</td>
<td>0.095</td>
<td>0.128</td>
<td>0.15</td>
<td>0.103</td>
<td>0.129</td>
</tr>
<tr>
<td>144</td>
<td>0.22</td>
<td>0.128</td>
<td>0.454</td>
<td>0.199</td>
<td>0.343</td>
<td>0.174</td>
<td>0.299</td>
<td>0.281</td>
</tr>
<tr>
<td>148</td>
<td>0.347</td>
<td>0.405</td>
<td>0.213</td>
<td>0.167</td>
<td>0.202</td>
<td>0.324</td>
<td>0.201</td>
<td>0.258</td>
</tr>
<tr>
<td>152</td>
<td>0.127</td>
<td>0.209</td>
<td>0.092</td>
<td>0.141</td>
<td>0.105</td>
<td>0.203</td>
<td>0.181</td>
<td>0.124</td>
</tr>
<tr>
<td>156</td>
<td>0.025</td>
<td>0.095</td>
<td>0.023</td>
<td>0.072</td>
<td>0.024</td>
<td>0.064</td>
<td>0.029</td>
<td>0.051</td>
</tr>
<tr>
<td>160</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0.013</td>
<td>0.004</td>
<td>0.009</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>D13S31</strong></td>
<td>58</td>
<td>77</td>
<td>88*</td>
<td>154</td>
<td>649</td>
<td>228</td>
<td></td>
<td></td>
</tr>
<tr>
<td>212</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>216</td>
<td>0.207</td>
<td>0</td>
<td>0.028</td>
<td>0.299</td>
<td>0.126</td>
<td>0.013</td>
<td>0.093</td>
<td>0.125</td>
</tr>
<tr>
<td>220</td>
<td>0.017</td>
<td>0</td>
<td>0.341</td>
<td>0.143</td>
<td>0.079</td>
<td>0.018</td>
<td>0.088</td>
<td>0.148</td>
</tr>
<tr>
<td>224</td>
<td>0.06</td>
<td>0.019</td>
<td>0.119</td>
<td>0.14</td>
<td>0.067</td>
<td>0.029</td>
<td>0.078</td>
<td>0.085</td>
</tr>
<tr>
<td>228</td>
<td>0.267</td>
<td>0.357</td>
<td>0.142</td>
<td>0.224</td>
<td>0.337</td>
<td>0.281</td>
<td>0.245</td>
<td>0.261</td>
</tr>
<tr>
<td>232</td>
<td>0.302</td>
<td>0.383</td>
<td>0.153</td>
<td>0.146</td>
<td>0.259</td>
<td>0.447</td>
<td>0.324</td>
<td>0.199</td>
</tr>
<tr>
<td>236</td>
<td>0.103</td>
<td>0.182</td>
<td>0.108</td>
<td>0.042</td>
<td>0.094</td>
<td>0.156</td>
<td>0.118</td>
<td>0.131</td>
</tr>
<tr>
<td>240</td>
<td>0.043</td>
<td>0.058</td>
<td>0.102</td>
<td>0.003</td>
<td>0.035</td>
<td>0.057</td>
<td>0.054</td>
<td>0.051</td>
</tr>
<tr>
<td>244</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>D16S53</strong></td>
<td>56</td>
<td>76</td>
<td>87</td>
<td>150</td>
<td>644</td>
<td>226</td>
<td></td>
<td></td>
</tr>
<tr>
<td>264</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0.013</td>
<td>0.025</td>
<td>0.027</td>
<td>0.044</td>
<td>0.011</td>
</tr>
<tr>
<td>268</td>
<td>0.125</td>
<td>0.184</td>
<td>0.086</td>
<td>0.25</td>
<td>0.12</td>
<td>0.188</td>
<td>0.191</td>
<td>0.107</td>
</tr>
<tr>
<td>272</td>
<td>0.027</td>
<td>0.105</td>
<td>0.264</td>
<td>0.093</td>
<td>0.06</td>
<td>0.139</td>
<td>0.078</td>
<td>0.152</td>
</tr>
<tr>
<td>276</td>
<td>0.295</td>
<td>0.211</td>
<td>0.264</td>
<td>0.3</td>
<td>0.306</td>
<td>0.263</td>
<td>0.26</td>
<td>0.315</td>
</tr>
<tr>
<td>280</td>
<td>0.321</td>
<td>0.27</td>
<td>0.287</td>
<td>0.227</td>
<td>0.3</td>
<td>0.221</td>
<td>0.27</td>
<td>0.298</td>
</tr>
<tr>
<td>284</td>
<td>0.223</td>
<td>0.178</td>
<td>0.08</td>
<td>0.087</td>
<td>0.161</td>
<td>0.142</td>
<td>0.132</td>
<td>0.118</td>
</tr>
<tr>
<td>288</td>
<td>0.009</td>
<td>0.033</td>
<td>0.017</td>
<td>0.03</td>
<td>0.027</td>
<td>0.02</td>
<td>0.025</td>
<td>0</td>
</tr>
<tr>
<td>292</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D18S51</td>
<td>56</td>
<td>78</td>
<td>88</td>
<td>149</td>
<td>640</td>
<td>217</td>
<td>99</td>
<td>89</td>
</tr>
<tr>
<td>--------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>271</td>
<td>0</td>
<td>0</td>
<td>0.011</td>
<td>0.003</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>275</td>
<td>0.009</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0.005</td>
<td>0.01</td>
<td>0.011</td>
</tr>
<tr>
<td>277</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>279</td>
<td>0.009</td>
<td>0.019</td>
<td>0</td>
<td>0.003</td>
<td>0.009</td>
<td>0.007</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>283</td>
<td>0.152</td>
<td>0.045</td>
<td>0.102</td>
<td>0.03</td>
<td>0.152</td>
<td>0.065</td>
<td>0.121</td>
<td>0.152</td>
</tr>
<tr>
<td>287</td>
<td>0.188</td>
<td>0.064</td>
<td>0.136</td>
<td>0.151</td>
<td>0.114</td>
<td>0.044</td>
<td>0.106</td>
<td>0.146</td>
</tr>
<tr>
<td>289</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>291</td>
<td>0.143</td>
<td>0.064</td>
<td>0.193</td>
<td>0.225</td>
<td>0.165</td>
<td>0.065</td>
<td>0.146</td>
<td>0.14</td>
</tr>
<tr>
<td>295</td>
<td>0.134</td>
<td>0.147</td>
<td>0.165</td>
<td>0.215</td>
<td>0.138</td>
<td>0.152</td>
<td>0.177</td>
<td>0.146</td>
</tr>
<tr>
<td>299</td>
<td>0.107</td>
<td>0.147</td>
<td>0.04</td>
<td>0.121</td>
<td>0.136</td>
<td>0.168</td>
<td>0.157</td>
<td>0.09</td>
</tr>
<tr>
<td>303</td>
<td>0.143</td>
<td>0.212</td>
<td>0.176</td>
<td>0.087</td>
<td>0.128</td>
<td>0.166</td>
<td>0.121</td>
<td>0.112</td>
</tr>
<tr>
<td>307</td>
<td>0.08</td>
<td>0.167</td>
<td>0.108</td>
<td>0.054</td>
<td>0.075</td>
<td>0.115</td>
<td>0.061</td>
<td>0.107</td>
</tr>
<tr>
<td>311</td>
<td>0</td>
<td>0.058</td>
<td>0.006</td>
<td>0.054</td>
<td>0.038</td>
<td>0.108</td>
<td>0.056</td>
<td>0.039</td>
</tr>
<tr>
<td>315</td>
<td>0.027</td>
<td>0.045</td>
<td>0.017</td>
<td>0.017</td>
<td>0.023</td>
<td>0.062</td>
<td>0.035</td>
<td>0.034</td>
</tr>
<tr>
<td>319</td>
<td>0</td>
<td>0.006</td>
<td>0.017</td>
<td>0.017</td>
<td>0.01</td>
<td>0.025</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>321</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>323</td>
<td>0</td>
<td>0.019</td>
<td>0.017</td>
<td>0.01</td>
<td>0.002</td>
<td>0.007</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>327</td>
<td>0</td>
<td>0</td>
<td>0.011</td>
<td>0.007</td>
<td>0.003</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>331</td>
<td>0.009</td>
<td>0.006</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>339</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D19S43</th>
<th>59</th>
<th>77</th>
<th>87</th>
<th>152</th>
<th>647</th>
<th>226</th>
<th>102</th>
<th>88</th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>105</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>109</td>
<td>0.017</td>
<td>0.071</td>
<td>0</td>
<td>0.003</td>
<td>0.004</td>
<td>0.055</td>
<td>0.025</td>
<td>0</td>
</tr>
<tr>
<td>111</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>113</td>
<td>0.119</td>
<td>0.162</td>
<td>0</td>
<td>0.039</td>
<td>0.096</td>
<td>0.131</td>
<td>0.108</td>
<td>0.057</td>
</tr>
<tr>
<td>115</td>
<td>0</td>
<td>0.058</td>
<td>0.017</td>
<td>0.007</td>
<td>0.003</td>
<td>0.033</td>
<td>0.015</td>
<td>0.011</td>
</tr>
<tr>
<td>117</td>
<td>0.297</td>
<td>0.299</td>
<td>0.092</td>
<td>0.247</td>
<td>0.218</td>
<td>0.274</td>
<td>0.206</td>
<td>0.148</td>
</tr>
<tr>
<td>119</td>
<td>0.034</td>
<td>0.032</td>
<td>0.167</td>
<td>0.033</td>
<td>0.009</td>
<td>0.038</td>
<td>0.02</td>
<td>0.097</td>
</tr>
<tr>
<td>121</td>
<td>0.28</td>
<td>0.169</td>
<td>0.259</td>
<td>0.319</td>
<td>0.355</td>
<td>0.221</td>
<td>0.309</td>
<td>0.301</td>
</tr>
<tr>
<td>123</td>
<td>0.017</td>
<td>0.065</td>
<td>0.046</td>
<td>0.122</td>
<td>0.019</td>
<td>0.053</td>
<td>0.054</td>
<td>0.045</td>
</tr>
<tr>
<td>125</td>
<td>0.153</td>
<td>0.071</td>
<td>0.161</td>
<td>0.039</td>
<td>0.173</td>
<td>0.066</td>
<td>0.142</td>
<td>0.193</td>
</tr>
<tr>
<td>127</td>
<td>0.042</td>
<td>0.019</td>
<td>0.19</td>
<td>0.155</td>
<td>0.055</td>
<td>0.06</td>
<td>0.078</td>
<td>0.074</td>
</tr>
<tr>
<td>129</td>
<td>0.025</td>
<td>0</td>
<td>0.023</td>
<td>0.01</td>
<td>0.045</td>
<td>0.029</td>
<td>0.025</td>
<td>0.04</td>
</tr>
<tr>
<td>131</td>
<td>0.008</td>
<td>0.019</td>
<td>0.04</td>
<td>0.023</td>
<td>0.018</td>
<td>0.029</td>
<td>0.02</td>
<td>0.023</td>
</tr>
<tr>
<td>133</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0.003</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>135</td>
<td>0.008</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>139</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>59</td>
<td>78</td>
<td>88</td>
<td>150</td>
<td>648</td>
<td>226</td>
<td>101</td>
<td>89</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>186</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>192</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>194</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>196</td>
<td>0</td>
<td>0.032</td>
<td>0.006</td>
<td>0.01</td>
<td>0.044</td>
<td>0.077</td>
<td>0.01</td>
<td>0.017</td>
</tr>
<tr>
<td>200</td>
<td>0.161</td>
<td>0.282</td>
<td>0.017</td>
<td>0.053</td>
<td>0.164</td>
<td>0.237</td>
<td>0.139</td>
<td>0.09</td>
</tr>
<tr>
<td>202</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>204</td>
<td>0.254</td>
<td>0.212</td>
<td>0.222</td>
<td>0.29</td>
<td>0.201</td>
<td>0.215</td>
<td>0.233</td>
<td>0.264</td>
</tr>
<tr>
<td>206</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>208</td>
<td>0.237</td>
<td>0.154</td>
<td>0.205</td>
<td>0.217</td>
<td>0.251</td>
<td>0.144</td>
<td>0.193</td>
<td>0.253</td>
</tr>
<tr>
<td>210</td>
<td>0.017</td>
<td>0.006</td>
<td>0.011</td>
<td>0.01</td>
<td>0.037</td>
<td>0.024</td>
<td>0.079</td>
<td>0.017</td>
</tr>
<tr>
<td>211</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>212</td>
<td>0.076</td>
<td>0.083</td>
<td>0.148</td>
<td>0.09</td>
<td>0.072</td>
<td>0.071</td>
<td>0.094</td>
<td>0.079</td>
</tr>
<tr>
<td>214</td>
<td>0.093</td>
<td>0.026</td>
<td>0.142</td>
<td>0.07</td>
<td>0.098</td>
<td>0.049</td>
<td>0.149</td>
<td>0.129</td>
</tr>
<tr>
<td>216</td>
<td>0.008</td>
<td>0</td>
<td>0.023</td>
<td>0.03</td>
<td>0.015</td>
<td>0.018</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>218</td>
<td>0.093</td>
<td>0.096</td>
<td>0.148</td>
<td>0.15</td>
<td>0.087</td>
<td>0.066</td>
<td>0.064</td>
<td>0.101</td>
</tr>
<tr>
<td>220</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0.001</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>222</td>
<td>0.059</td>
<td>0.045</td>
<td>0.063</td>
<td>0.07</td>
<td>0.02</td>
<td>0.033</td>
<td>0.03</td>
<td>0.022</td>
</tr>
<tr>
<td>224</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.015</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>226</td>
<td>0</td>
<td>0.011</td>
<td>0.003</td>
<td>0.005</td>
<td>0.002</td>
<td>0</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>228</td>
<td>0</td>
<td>0.038</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.033</td>
<td>0.01</td>
<td>0.006</td>
</tr>
<tr>
<td>230</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>232</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>234</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FGA</td>
<td>58</td>
<td>77</td>
<td>87</td>
<td>151</td>
<td>643</td>
<td>226</td>
<td>102*</td>
<td>89</td>
</tr>
<tr>
<td>-----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>----</td>
</tr>
<tr>
<td>206</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>210</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>214</td>
<td>0.009</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>218</td>
<td>0.009</td>
<td>0.006</td>
<td>0.006</td>
<td>0.03</td>
<td>0.014</td>
<td>0.004</td>
<td>0.01</td>
<td>0.017</td>
</tr>
<tr>
<td>220</td>
<td>0</td>
<td>0.019</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>222</td>
<td>0.069</td>
<td>0.052</td>
<td>0.172</td>
<td>0.076</td>
<td>0.071</td>
<td>0.066</td>
<td>0.059</td>
<td>0.101</td>
</tr>
<tr>
<td>224</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>226</td>
<td>0.121</td>
<td>0.065</td>
<td>0.063</td>
<td>0.07</td>
<td>0.135</td>
<td>0.035</td>
<td>0.118</td>
<td>0.096</td>
</tr>
<tr>
<td>228</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>230</td>
<td>0.19</td>
<td>0.11</td>
<td>0.109</td>
<td>0.123</td>
<td>0.15</td>
<td>0.133</td>
<td>0.162</td>
<td>0.152</td>
</tr>
<tr>
<td>232</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.013</td>
<td>0.002</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td>234</td>
<td>0.155</td>
<td>0.169</td>
<td>0.121</td>
<td>0.195</td>
<td>0.198</td>
<td>0.181</td>
<td>0.157</td>
<td>0.112</td>
</tr>
<tr>
<td>236</td>
<td>0.017</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.012</td>
<td>0.002</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>238</td>
<td>0.129</td>
<td>0.253</td>
<td>0.086</td>
<td>0.169</td>
<td>0.152</td>
<td>0.146</td>
<td>0.147</td>
<td>0.129</td>
</tr>
<tr>
<td>240</td>
<td>0.009</td>
<td>0</td>
<td>0.011</td>
<td>0.013</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>242</td>
<td>0.155</td>
<td>0.123</td>
<td>0.138</td>
<td>0.162</td>
<td>0.138</td>
<td>0.173</td>
<td>0.147</td>
<td>0.18</td>
</tr>
<tr>
<td>244</td>
<td>0.009</td>
<td>0</td>
<td>0.006</td>
<td>0.01</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>245</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>246</td>
<td>0.095</td>
<td>0.078</td>
<td>0.155</td>
<td>0.093</td>
<td>0.075</td>
<td>0.104</td>
<td>0.098</td>
<td>0.135</td>
</tr>
<tr>
<td>250</td>
<td>0.026</td>
<td>0.032</td>
<td>0.115</td>
<td>0.04</td>
<td>0.03</td>
<td>0.049</td>
<td>0.054</td>
<td>0.039</td>
</tr>
<tr>
<td>254</td>
<td>0.009</td>
<td>0.032</td>
<td>0.011</td>
<td>0.003</td>
<td>0.007</td>
<td>0.029</td>
<td>0.039</td>
<td>0.028</td>
</tr>
<tr>
<td>258</td>
<td>0</td>
<td>0.045</td>
<td>0.006</td>
<td>0</td>
<td>0.002</td>
<td>0.033</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>262</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.004</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>266</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>268</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TH01</th>
<th>59</th>
<th>73</th>
<th>87</th>
<th>149</th>
<th>645</th>
<th>226</th>
<th>102</th>
<th>89</th>
</tr>
</thead>
<tbody>
<tr>
<td>167</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>171</td>
<td>0.212</td>
<td>0.075</td>
<td>0.299</td>
<td>0.084</td>
<td>0.236</td>
<td>0.164</td>
<td>0.201</td>
<td>0.236</td>
</tr>
<tr>
<td>175</td>
<td>0.161</td>
<td>0.37</td>
<td>0.443</td>
<td>0.332</td>
<td>0.173</td>
<td>0.389</td>
<td>0.25</td>
<td>0.32</td>
</tr>
<tr>
<td>179</td>
<td>0.102</td>
<td>0.288</td>
<td>0.006</td>
<td>0.07</td>
<td>0.099</td>
<td>0.219</td>
<td>0.132</td>
<td>0.112</td>
</tr>
<tr>
<td>182</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>183</td>
<td>0.186</td>
<td>0.212</td>
<td>0.011</td>
<td>0.433</td>
<td>0.157</td>
<td>0.144</td>
<td>0.216</td>
<td>0.107</td>
</tr>
<tr>
<td>186</td>
<td>0.331</td>
<td>0.055</td>
<td>0.241</td>
<td>0.03</td>
<td>0.317</td>
<td>0.071</td>
<td>0.201</td>
<td>0.219</td>
</tr>
<tr>
<td>187</td>
<td>0.008</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0.014</td>
<td>0.007</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>191</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TPOX</td>
<td>59</td>
<td>75</td>
<td>87</td>
<td>151</td>
<td>646</td>
<td>223</td>
<td>102</td>
<td>88</td>
</tr>
<tr>
<td>------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>218</td>
<td>0.008</td>
<td>0.002</td>
<td>0.004</td>
<td>0.004</td>
<td>0.006</td>
<td>0.015</td>
<td>0.006</td>
<td>0.004</td>
</tr>
<tr>
<td>222</td>
<td>0.009</td>
<td>0.002</td>
<td>0.004</td>
<td>0.004</td>
<td>0.006</td>
<td>0.015</td>
<td>0.006</td>
<td>0.004</td>
</tr>
<tr>
<td>226</td>
<td>0.066</td>
<td>0.033</td>
<td>0.062</td>
<td>0.054</td>
<td>0.035</td>
<td>0.045</td>
<td>0.046</td>
<td>0.036</td>
</tr>
<tr>
<td>234</td>
<td>0.144</td>
<td>0.023</td>
<td>0.093</td>
<td>0.104</td>
<td>0.206</td>
<td>0.088</td>
<td>0.108</td>
<td>0.088</td>
</tr>
<tr>
<td>238</td>
<td>0.059</td>
<td>0.12</td>
<td>0.017</td>
<td>0.007</td>
<td>0.053</td>
<td>0.092</td>
<td>0.039</td>
<td>0.057</td>
</tr>
<tr>
<td>242</td>
<td>0.297</td>
<td>0.236</td>
<td>0.248</td>
<td>0.261</td>
<td>0.224</td>
<td>0.304</td>
<td>0.278</td>
<td>0.278</td>
</tr>
<tr>
<td>246</td>
<td>0.025</td>
<td>0.013</td>
<td>0.023</td>
<td>0.032</td>
<td>0.029</td>
<td>0.083</td>
<td>0.088</td>
<td>0.088</td>
</tr>
<tr>
<td>250</td>
<td>0.006</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.005</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>254</td>
<td>0.066</td>
<td>0.033</td>
<td>0.062</td>
<td>0.054</td>
<td>0.035</td>
<td>0.045</td>
<td>0.046</td>
<td>0.036</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>vWA</th>
<th>58</th>
<th>76</th>
<th>87</th>
<th>150</th>
<th>649</th>
<th>227</th>
<th>102</th>
<th>88</th>
</tr>
</thead>
<tbody>
<tr>
<td>154</td>
<td>0.013</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>158</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>162</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>166</td>
<td>0.103</td>
<td>0.059</td>
<td>0.04</td>
<td>0.303</td>
<td>0.11</td>
<td>0.068</td>
<td>0.093</td>
<td>0.068</td>
</tr>
<tr>
<td>170</td>
<td>0.155</td>
<td>0.151</td>
<td>0.034</td>
<td>0.037</td>
<td>0.102</td>
<td>0.26</td>
<td>0.098</td>
<td>0.108</td>
</tr>
<tr>
<td>174</td>
<td>0.233</td>
<td>0.23</td>
<td>0.466</td>
<td>0.113</td>
<td>0.213</td>
<td>0.244</td>
<td>0.225</td>
<td>0.273</td>
</tr>
<tr>
<td>178</td>
<td>0.259</td>
<td>0.224</td>
<td>0.316</td>
<td>0.19</td>
<td>0.267</td>
<td>0.185</td>
<td>0.294</td>
<td>0.301</td>
</tr>
<tr>
<td>182</td>
<td>0.164</td>
<td>0.118</td>
<td>0.103</td>
<td>0.217</td>
<td>0.206</td>
<td>0.13</td>
<td>0.196</td>
<td>0.21</td>
</tr>
<tr>
<td>186</td>
<td>0.078</td>
<td>0.132</td>
<td>0.034</td>
<td>0.127</td>
<td>0.085</td>
<td>0.064</td>
<td>0.054</td>
<td>0.04</td>
</tr>
<tr>
<td>190</td>
<td>0.009</td>
<td>0.053</td>
<td>0.013</td>
<td>0.014</td>
<td>0.018</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>194</td>
<td>0.006</td>
<td>0.000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>226</td>
<td>0.002</td>
<td>0.000</td>
<td>0.004</td>
<td>0.004</td>
<td>0.022</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>230</td>
<td>0.466</td>
<td>0.333</td>
<td>0.506</td>
<td>0.629</td>
<td>0.543</td>
<td>0.365</td>
<td>0.451</td>
<td>0.466</td>
</tr>
<tr>
<td>234</td>
<td>0.144</td>
<td>0.18</td>
<td>0.023</td>
<td>0.093</td>
<td>0.104</td>
<td>0.206</td>
<td>0.088</td>
<td>0.108</td>
</tr>
<tr>
<td>238</td>
<td>0.059</td>
<td>0.12</td>
<td>0.017</td>
<td>0.007</td>
<td>0.053</td>
<td>0.092</td>
<td>0.039</td>
<td>0.057</td>
</tr>
<tr>
<td>242</td>
<td>0.297</td>
<td>0.236</td>
<td>0.248</td>
<td>0.261</td>
<td>0.224</td>
<td>0.304</td>
<td>0.278</td>
<td>0.278</td>
</tr>
<tr>
<td>246</td>
<td>0.025</td>
<td>0.013</td>
<td>0.023</td>
<td>0.032</td>
<td>0.029</td>
<td>0.083</td>
<td>0.088</td>
<td>0.088</td>
</tr>
<tr>
<td>250</td>
<td>0.006</td>
<td>0.000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>254</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>D2S319</td>
<td></td>
<td></td>
<td></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>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>642</td>
<td>222</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>120</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>0.001</td>
<td>0.007</td>
<td>0</td>
<td>na</td>
</tr>
<tr>
<td>126</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>0.13</td>
<td>0.135</td>
<td>0.147</td>
<td>na</td>
</tr>
<tr>
<td>128</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>0.038</td>
<td>0.083</td>
<td>0.066</td>
<td>na</td>
</tr>
<tr>
<td>130</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>0.342</td>
<td>0.275</td>
<td>0.368</td>
<td>na</td>
</tr>
<tr>
<td>132</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>0.271</td>
<td>0.104</td>
<td>0.191</td>
<td>na</td>
</tr>
<tr>
<td>134</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>0.208</td>
<td>0.36</td>
<td>0.228</td>
<td>na</td>
</tr>
<tr>
<td>136</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>0.01</td>
<td>0.036</td>
<td>0</td>
<td>na</td>
</tr>
<tr>
<td></td>
<td>D17S79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>59</td>
<td>78</td>
<td>88</td>
<td>154</td>
<td>636</td>
<td>213</td>
<td>101</td>
<td>88</td>
</tr>
<tr>
<td>185</td>
<td>0.008</td>
<td>0.019</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.007</td>
<td>0.005</td>
<td>0</td>
</tr>
<tr>
<td>187</td>
<td>0.008</td>
<td>0.109</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
<td>0.15</td>
<td>0.045</td>
<td>0.023</td>
</tr>
<tr>
<td>189</td>
<td>0.398</td>
<td>0.09</td>
<td>0.023</td>
<td>0.013</td>
<td>0.353</td>
<td>0.167</td>
<td>0.287</td>
<td>0.239</td>
</tr>
<tr>
<td>191</td>
<td>0.017</td>
<td>0.032</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0.031</td>
<td>0.045</td>
<td>0.006</td>
</tr>
<tr>
<td>193</td>
<td>0.042</td>
<td>0.34</td>
<td>0.017</td>
<td>0.068</td>
<td>0.048</td>
<td>0.218</td>
<td>0.099</td>
<td>0.074</td>
</tr>
<tr>
<td>195</td>
<td>0.186</td>
<td>0.103</td>
<td>0.148</td>
<td>0.175</td>
<td>0.257</td>
<td>0.11</td>
<td>0.134</td>
<td>0.216</td>
</tr>
<tr>
<td>197</td>
<td>0.025</td>
<td>0.109</td>
<td>0.023</td>
<td>0.045</td>
<td>0.029</td>
<td>0.113</td>
<td>0.035</td>
<td>0.034</td>
</tr>
<tr>
<td>199</td>
<td>0.22</td>
<td>0.154</td>
<td>0.761</td>
<td>0.5</td>
<td>0.185</td>
<td>0.153</td>
<td>0.272</td>
<td>0.335</td>
</tr>
<tr>
<td>201</td>
<td>0.017</td>
<td>0.032</td>
<td>0</td>
<td>0.045</td>
<td>0.024</td>
<td>0.026</td>
<td>0.035</td>
<td>0.04</td>
</tr>
<tr>
<td>203</td>
<td>0.076</td>
<td>0.013</td>
<td>0.028</td>
<td>0.149</td>
<td>0.087</td>
<td>0.026</td>
<td>0.045</td>
<td>0.034</td>
</tr>
<tr>
<td>205</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>207</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D8S272</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>49</td>
<td>74</td>
<td>88</td>
<td>148</td>
<td>629</td>
<td>217</td>
<td>97</td>
<td>84**</td>
</tr>
<tr>
<td>235</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>237</td>
<td>0.02</td>
<td>0.074</td>
<td>0.148</td>
<td>0.088</td>
<td>0.034</td>
<td>0.048</td>
<td>0.036</td>
<td>0.077</td>
</tr>
<tr>
<td>239</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.001</td>
<td>0.002</td>
<td>0.005</td>
<td>0</td>
</tr>
<tr>
<td>241</td>
<td>0.092</td>
<td>0.034</td>
<td>0.006</td>
<td>0.037</td>
<td>0.122</td>
<td>0.051</td>
<td>0.015</td>
<td>0.083</td>
</tr>
<tr>
<td>243</td>
<td>0</td>
<td>0.054</td>
<td>0.011</td>
<td>0.01</td>
<td>0.01</td>
<td>0.025</td>
<td>0.026</td>
<td>0.03</td>
</tr>
<tr>
<td>245</td>
<td>0.388</td>
<td>0.034</td>
<td>0.222</td>
<td>0.355</td>
<td>0.398</td>
<td>0.134</td>
<td>0.361</td>
<td>0.274</td>
</tr>
<tr>
<td>247</td>
<td>0.082</td>
<td>0.236</td>
<td>0.023</td>
<td>0.253</td>
<td>0.054</td>
<td>0.249</td>
<td>0.17</td>
<td>0.083</td>
</tr>
<tr>
<td>249</td>
<td>0.031</td>
<td>0.047</td>
<td>0.119</td>
<td>0.047</td>
<td>0.01</td>
<td>0.065</td>
<td>0.052</td>
<td>0.042</td>
</tr>
<tr>
<td>251</td>
<td>0.214</td>
<td>0.196</td>
<td>0.415</td>
<td>0.182</td>
<td>0.184</td>
<td>0.157</td>
<td>0.186</td>
<td>0.232</td>
</tr>
<tr>
<td>253</td>
<td>0.041</td>
<td>0.122</td>
<td>0.04</td>
<td>0.014</td>
<td>0.027</td>
<td>0.108</td>
<td>0.052</td>
<td>0.042</td>
</tr>
<tr>
<td>254</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>255</td>
<td>0.071</td>
<td>0.169</td>
<td>0.017</td>
<td>0</td>
<td>0.111</td>
<td>0.118</td>
<td>0.041</td>
<td>0.077</td>
</tr>
<tr>
<td>257</td>
<td>0.051</td>
<td>0.034</td>
<td>0</td>
<td>0.007</td>
<td>0.04</td>
<td>0.035</td>
<td>0.046</td>
<td>0.054</td>
</tr>
<tr>
<td>259</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.006</td>
<td>0.002</td>
<td>0.005</td>
<td>0</td>
</tr>
<tr>
<td>261</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D1S196</td>
<td>59</td>
<td>78</td>
<td>87</td>
<td>151</td>
<td>542</td>
<td>220</td>
<td>100</td>
<td>89</td>
</tr>
<tr>
<td>--------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>320</td>
<td>0.39</td>
<td>0.34</td>
<td>0.425</td>
<td>0.656</td>
<td>0.351</td>
<td>0.339</td>
<td>0.275</td>
<td>0.331</td>
</tr>
<tr>
<td>322</td>
<td>0.025</td>
<td>0</td>
<td>0.017</td>
<td>0.01</td>
<td>0.018</td>
<td>0.016</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>324</td>
<td>0.169</td>
<td>0.019</td>
<td>0.253</td>
<td>0.96</td>
<td>0.241</td>
<td>0.075</td>
<td>0.205</td>
<td>0.242</td>
</tr>
<tr>
<td>326</td>
<td>0.102</td>
<td>0.16</td>
<td>0.04</td>
<td>0.046</td>
<td>0.072</td>
<td>0.093</td>
<td>0.095</td>
<td>0.073</td>
</tr>
<tr>
<td>330</td>
<td>0.212</td>
<td>0.346</td>
<td>0.253</td>
<td>0.185</td>
<td>0.216</td>
<td>0.343</td>
<td>0.275</td>
<td>0.242</td>
</tr>
<tr>
<td>332</td>
<td>0.102</td>
<td>0.058</td>
<td>0.111</td>
<td>0.097</td>
<td>0.08</td>
<td>0.125</td>
<td>0.096</td>
<td></td>
</tr>
<tr>
<td>334</td>
<td>0</td>
<td>0.077</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0.041</td>
<td>0.02</td>
<td>0.006</td>
</tr>
<tr>
<td>336</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

| D7S640 | 53 | 63 | 65 | 96* | 615 | 200 | 95 | 65* |
|--------|----|----|----|-----|-----|-----|----|
| 104    | 0 | 0 | 0 | 0.016 | 0.001 | 0 | 0.016 | 0 |
| 106    | 0 | 0.167 | 0 | 0.005 | 0.001 | 0.065 | 0.016 | 0 |
| 108    | 0 | 0.016 | 0 | 0 | 0.002 | 0.018 | 0 | 0.008 |
| 110    | 0.028 | 0.048 | 0 | 0 | 0.001 | 0 | 0 | 0.008 |
| 112    | 0.009 | 0.103 | 0 | 0.01 | 0.008 | 0.135 | 0.053 | 0.023 |
| 114    | 0.038 | 0 | 0.031 | 0.01 | 0.033 | 0.013 | 0.021 | 0.015 |
| 116    | 0.217 | 0.04 | 0.177 | 0.302 | 0.189 | 0.113 | 0.168 | 0.238 |
| 118    | 0.104 | 0.04 | 0.023 | 0.104 | 0.063 | 0.05 | 0.084 | 0.069 |
| 120    | 0.179 | 0.048 | 0.031 | 0.094 | 0.174 | 0.058 | 0.132 | 0.092 |
| 122    | 0.16 | 0.103 | 0.138 | 0.156 | 0.183 | 0.135 | 0.1 | 0.185 |
| 124    | 0.113 | 0.079 | 0.023 | 0.099 | 0.123 | 0.073 | 0.137 | 0.1 |
| 126    | 0.047 | 0.063 | 0.023 | 0.052 | 0.081 | 0.083 | 0.053 | 0.038 |
| 128    | 0.057 | 0.135 | 0.077 | 0.057 | 0.045 | 0.063 | 0.068 | 0.046 |
| 130    | 0 | 0.063 | 0.008 | 0.063 | 0.02 | 0.05 | 0.026 | 0.038 |
| 132    | 0.009 | 0.063 | 0.1 | 0.031 | 0.019 | 0.055 | 0.021 | 0.038 |
| 134    | 0 | 0.024 | 0.238 | 0 | 0.01 | 0.053 | 0.047 | 0.046 |
| 136    | 0.028 | 0 | 0.054 | 0 | 0.015 | 0.018 | 0.026 | 0.031 |
| 138    | 0.009 | 0.008 | 0.023 | 0 | 0.015 | 0.015 | 0.005 | 0.023 |
| 140    | 0 | 0 | 0.046 | 0 | 0.008 | 0.005 | 0.021 | 0 |
| 142    | 0 | 0 | 0.008 | 0 | 0.005 | 0.003 | 0.005 | 0 |
| 144    | 0 | 0 | 0 | 0 | 0.003 | 0 | 0 | 0 |
| 146    | 0 | 0 | 0 | 0 | 0.001 | 0 | 0 | 0 |

<table>
<thead>
<tr>
<th>D8S182</th>
<th>59**</th>
<th>78</th>
<th>88</th>
<th>154</th>
<th>643</th>
<th>226</th>
<th>101</th>
<th>88</th>
</tr>
</thead>
<tbody>
<tr>
<td>154</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>156</td>
<td>0.068</td>
<td>0.019</td>
<td>0.153</td>
<td>0</td>
<td>0.053</td>
<td>0.027</td>
<td>0.059</td>
<td>0.068</td>
</tr>
<tr>
<td>158</td>
<td>0.034</td>
<td>0.378</td>
<td>0.006</td>
<td>0.032</td>
<td>0.031</td>
<td>0.33</td>
<td>0.109</td>
<td>0.057</td>
</tr>
<tr>
<td>160</td>
<td>0.475</td>
<td>0.173</td>
<td>0.511</td>
<td>0.558</td>
<td>0.541</td>
<td>0.23</td>
<td>0.47</td>
<td>0.528</td>
</tr>
<tr>
<td>162</td>
<td>0.076</td>
<td>0.103</td>
<td>0.006</td>
<td>0.078</td>
<td>0.034</td>
<td>0.075</td>
<td>0.054</td>
<td>0.023</td>
</tr>
<tr>
<td>164</td>
<td>0.322</td>
<td>0.263</td>
<td>0.324</td>
<td>0.321</td>
<td>0.31</td>
<td>0.263</td>
<td>0.282</td>
<td>0.295</td>
</tr>
<tr>
<td>166</td>
<td>0.025</td>
<td>0.051</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
<td>0.073</td>
<td>0.02</td>
<td>0.023</td>
</tr>
<tr>
<td>168</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>66</td>
<td>82</td>
<td>106</td>
<td>623**</td>
<td>214</td>
<td>95</td>
<td>77</td>
</tr>
<tr>
<td>------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-------</td>
<td>-----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>246</td>
<td>0</td>
<td>0.076</td>
<td>0</td>
<td>0.005</td>
<td>0.001</td>
<td>0.075</td>
<td>0.016</td>
<td>0.006</td>
</tr>
<tr>
<td>248</td>
<td>0.079</td>
<td>0.174</td>
<td>0.018</td>
<td>0.156</td>
<td>0.059</td>
<td>0.131</td>
<td>0.058</td>
<td>0.045</td>
</tr>
<tr>
<td>250</td>
<td>0.039</td>
<td>0.462</td>
<td>0.28</td>
<td>0.387</td>
<td>0.055</td>
<td>0.376</td>
<td>0.179</td>
<td>0.201</td>
</tr>
<tr>
<td>252</td>
<td>0.184</td>
<td>0.265</td>
<td>0.183</td>
<td>0.17</td>
<td>0.185</td>
<td>0.241</td>
<td>0.237</td>
<td>0.253</td>
</tr>
<tr>
<td>254</td>
<td>0.066</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.039</td>
<td>0.035</td>
<td>0.037</td>
<td>0</td>
</tr>
<tr>
<td>256</td>
<td>0</td>
<td>0.008</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0.019</td>
<td>0.011</td>
<td>0</td>
</tr>
<tr>
<td>258</td>
<td>0.066</td>
<td>0</td>
<td>0.018</td>
<td>0.009</td>
<td>0.045</td>
<td>0.012</td>
<td>0.042</td>
<td>0.052</td>
</tr>
<tr>
<td>260</td>
<td>0.092</td>
<td>0.008</td>
<td>0.183</td>
<td>0.042</td>
<td>0.1</td>
<td>0.035</td>
<td>0.105</td>
<td>0.117</td>
</tr>
<tr>
<td>262</td>
<td>0.395</td>
<td>0</td>
<td>0.177</td>
<td>0.189</td>
<td>0.408</td>
<td>0.056</td>
<td>0.253</td>
<td>0.24</td>
</tr>
<tr>
<td>264</td>
<td>0.066</td>
<td>0.008</td>
<td>0.079</td>
<td>0.042</td>
<td>0.089</td>
<td>0.016</td>
<td>0.053</td>
<td>0.071</td>
</tr>
<tr>
<td>266</td>
<td>0.013</td>
<td>0</td>
<td>0.055</td>
<td>0</td>
<td>0.007</td>
<td>0.005</td>
<td>0.011</td>
<td>0.013</td>
</tr>
<tr>
<td>268</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>270</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>272</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>274</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>59</th>
<th>77</th>
<th>87</th>
<th>154</th>
<th>642</th>
<th>218</th>
<th>100</th>
<th>88</th>
</tr>
</thead>
<tbody>
<tr>
<td>272</td>
<td>0.008</td>
<td>0.052</td>
<td>0.029</td>
<td>0.237</td>
<td>0.002</td>
<td>0.089</td>
<td>0.02</td>
<td>0.028</td>
</tr>
<tr>
<td>274</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.019</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>276</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>278</td>
<td>0</td>
<td>0.032</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0</td>
<td>0.03</td>
<td>0</td>
</tr>
<tr>
<td>280</td>
<td>0.102</td>
<td>0.143</td>
<td>0.04</td>
<td>0.016</td>
<td>0.102</td>
<td>0.126</td>
<td>0.105</td>
<td>0.131</td>
</tr>
<tr>
<td>282</td>
<td>0.347</td>
<td>0.045</td>
<td>0.109</td>
<td>0.227</td>
<td>0.275</td>
<td>0.112</td>
<td>0.14</td>
<td>0.199</td>
</tr>
<tr>
<td>284</td>
<td>0.22</td>
<td>0.377</td>
<td>0.287</td>
<td>0.127</td>
<td>0.303</td>
<td>0.335</td>
<td>0.365</td>
<td>0.29</td>
</tr>
<tr>
<td>286</td>
<td>0.102</td>
<td>0.065</td>
<td>0.236</td>
<td>0.075</td>
<td>0.095</td>
<td>0.096</td>
<td>0.11</td>
<td>0.142</td>
</tr>
<tr>
<td>288</td>
<td>0.102</td>
<td>0.11</td>
<td>0.253</td>
<td>0.198</td>
<td>0.104</td>
<td>0.103</td>
<td>0.14</td>
<td>0.114</td>
</tr>
<tr>
<td>290</td>
<td>0.119</td>
<td>0.143</td>
<td>0.046</td>
<td>0.078</td>
<td>0.098</td>
<td>0.119</td>
<td>0.075</td>
<td>0.08</td>
</tr>
<tr>
<td>292</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0.016</td>
<td>0.009</td>
<td>0.005</td>
<td>0.015</td>
<td>0.017</td>
</tr>
<tr>
<td>294</td>
<td>0</td>
<td>0.019</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.014</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>296</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>53</td>
<td>58</td>
<td>83</td>
<td>128</td>
<td>522</td>
<td>158</td>
<td>85</td>
<td>70</td>
</tr>
<tr>
<td>--------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>D14S68</td>
<td>308</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>310</td>
<td>0</td>
<td>0.103</td>
<td>0</td>
<td>0</td>
<td>0.019</td>
<td>0</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>312</td>
<td>0</td>
<td>0.078</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0.006</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>314</td>
<td>0</td>
<td>0.052</td>
<td>0</td>
<td>0.004</td>
<td>0.001</td>
<td>0.114</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>316</td>
<td>0.009</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0.016</td>
<td>0.019</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>318</td>
<td>0.142</td>
<td>0.078</td>
<td>0.181</td>
<td>0.008</td>
<td>0.128</td>
<td>0.079</td>
<td>0.159</td>
</tr>
<tr>
<td></td>
<td>320</td>
<td>0.075</td>
<td>0.069</td>
<td>0.036</td>
<td>0.008</td>
<td>0.049</td>
<td>0.117</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td>322</td>
<td>0.038</td>
<td>0.043</td>
<td>0.012</td>
<td>0.027</td>
<td>0.049</td>
<td>0.047</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>324</td>
<td>0.311</td>
<td>0.25</td>
<td>0.488</td>
<td>0.387</td>
<td>0.307</td>
<td>0.285</td>
<td>0.253</td>
</tr>
<tr>
<td></td>
<td>326</td>
<td>0.17</td>
<td>0.216</td>
<td>0.09</td>
<td>0.293</td>
<td>0.187</td>
<td>0.199</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>328</td>
<td>0.132</td>
<td>0.078</td>
<td>0.084</td>
<td>0.125</td>
<td>0.116</td>
<td>0.07</td>
<td>0.124</td>
</tr>
<tr>
<td></td>
<td>330</td>
<td>0.113</td>
<td>0.017</td>
<td>0.084</td>
<td>0.098</td>
<td>0.103</td>
<td>0.028</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td>332</td>
<td>0.009</td>
<td>0.017</td>
<td>0.018</td>
<td>0.047</td>
<td>0.035</td>
<td>0.016</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>334</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.004</td>
<td>0.006</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>336</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>58</th>
<th>78</th>
<th>88</th>
<th>153</th>
<th>575</th>
<th>225</th>
<th>102</th>
<th>89</th>
</tr>
</thead>
<tbody>
<tr>
<td>D5S407</td>
<td>84</td>
<td>0.19</td>
<td>0.051</td>
<td>0.148</td>
<td>0.278</td>
<td>0.157</td>
<td>0.033</td>
<td>0.152</td>
</tr>
<tr>
<td></td>
<td>86</td>
<td>0.026</td>
<td>0.045</td>
<td>0</td>
<td>0.007</td>
<td>0.017</td>
<td>0.062</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>88</td>
<td>0.112</td>
<td>0.237</td>
<td>0.216</td>
<td>0.141</td>
<td>0.121</td>
<td>0.242</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>0.043</td>
<td>0.045</td>
<td>0.011</td>
<td>0.023</td>
<td>0.035</td>
<td>0.024</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>92</td>
<td>0.009</td>
<td>0.109</td>
<td>0.006</td>
<td>0.007</td>
<td>0.03</td>
<td>0.1</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>94</td>
<td>0.095</td>
<td>0.071</td>
<td>0.006</td>
<td>0.016</td>
<td>0.096</td>
<td>0.062</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>0.06</td>
<td>0.064</td>
<td>0.006</td>
<td>0.029</td>
<td>0.054</td>
<td>0.062</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>98</td>
<td>0.103</td>
<td>0.128</td>
<td>0.176</td>
<td>0.092</td>
<td>0.158</td>
<td>0.129</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.198</td>
<td>0.122</td>
<td>0.074</td>
<td>0.16</td>
<td>0.208</td>
<td>0.138</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>0.155</td>
<td>0.026</td>
<td>0.176</td>
<td>0.114</td>
<td>0.098</td>
<td>0.042</td>
<td>0.137</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>0.009</td>
<td>0.096</td>
<td>0.051</td>
<td>0.095</td>
<td>0.017</td>
<td>0.069</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>106</td>
<td>0</td>
<td>0.006</td>
<td>0.097</td>
<td>0.023</td>
<td>0.007</td>
<td>0.016</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>108</td>
<td>0</td>
<td>0</td>
<td>0.034</td>
<td>0.013</td>
<td>0.003</td>
<td>0.016</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.004</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>112</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>59</th>
<th>78</th>
<th>88</th>
<th>154</th>
<th>625</th>
<th>219</th>
<th>100</th>
<th>87</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2S162</td>
<td>114</td>
<td>0</td>
<td>0.019</td>
<td>0</td>
<td>0</td>
<td>0.021</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>116</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
<td>0.185</td>
<td>0.001</td>
<td>0.005</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>118</td>
<td>0.008</td>
<td>0.006</td>
<td>0.006</td>
<td>0</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>0.051</td>
<td>0.186</td>
<td>0.034</td>
<td>0.055</td>
<td>0.109</td>
<td>0.114</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>122</td>
<td>0</td>
<td>0.026</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0.016</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>0</td>
<td>0.103</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.082</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>126</td>
<td>0</td>
<td>0.135</td>
<td>0</td>
<td>0.003</td>
<td>0.004</td>
<td>0.105</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>0.11</td>
<td>0.077</td>
<td>0.017</td>
<td>0.055</td>
<td>0.042</td>
<td>0.155</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>0.059</td>
<td>0.115</td>
<td>0.011</td>
<td>0.088</td>
<td>0.056</td>
<td>0.089</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>0.322</td>
<td>0.205</td>
<td>0.375</td>
<td>0.364</td>
<td>0.322</td>
<td>0.221</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D10S19</td>
<td>59</td>
<td>78</td>
<td>88</td>
<td>154</td>
<td>638</td>
<td>227</td>
<td>100</td>
<td>89</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>156</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>158</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>160</td>
<td>0</td>
<td>0.077</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.046</td>
<td>0.005</td>
<td>0.017</td>
</tr>
<tr>
<td>162</td>
<td>0</td>
<td>0.019</td>
<td>0</td>
<td>0.003</td>
<td>0.001</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>164</td>
<td>0.068</td>
<td>0.019</td>
<td>0</td>
<td>0</td>
<td>0.032</td>
<td>0.042</td>
<td>0.05</td>
<td>0.028</td>
</tr>
<tr>
<td>166</td>
<td>0</td>
<td>0.09</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.031</td>
<td>0.015</td>
<td>0.006</td>
</tr>
<tr>
<td>168</td>
<td>0.11</td>
<td>0.103</td>
<td>0.063</td>
<td>0.097</td>
<td>0.091</td>
<td>0.097</td>
<td>0.11</td>
<td>0.079</td>
</tr>
<tr>
<td>170</td>
<td>0.144</td>
<td>0.154</td>
<td>0.17</td>
<td>0.227</td>
<td>0.158</td>
<td>0.218</td>
<td>0.19</td>
<td>0.197</td>
</tr>
<tr>
<td>172</td>
<td>0.178</td>
<td>0.359</td>
<td>0.42</td>
<td>0.299</td>
<td>0.179</td>
<td>0.24</td>
<td>0.2</td>
<td>0.281</td>
</tr>
<tr>
<td>174</td>
<td>0.381</td>
<td>0.032</td>
<td>0.256</td>
<td>0.273</td>
<td>0.411</td>
<td>0.17</td>
<td>0.36</td>
<td>0.337</td>
</tr>
<tr>
<td>176</td>
<td>0.102</td>
<td>0.083</td>
<td>0.08</td>
<td>0.058</td>
<td>0.113</td>
<td>0.108</td>
<td>0.06</td>
<td>0.039</td>
</tr>
<tr>
<td>178</td>
<td>0.008</td>
<td>0.013</td>
<td>0</td>
<td>0.032</td>
<td>0.009</td>
<td>0.007</td>
<td>0</td>
<td>0.011</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D11S93</th>
<th>53</th>
<th>77</th>
<th>88</th>
<th>153</th>
<th>637**</th>
<th>221</th>
<th>100</th>
<th>89</th>
</tr>
</thead>
<tbody>
<tr>
<td>192</td>
<td>0</td>
<td>0</td>
<td>0.074</td>
<td>0.141</td>
<td>0.015</td>
<td>0</td>
<td>0.02</td>
<td>0.028</td>
</tr>
<tr>
<td>194</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>196</td>
<td>0</td>
<td>0.091</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.084</td>
<td>0.015</td>
<td>0.006</td>
</tr>
<tr>
<td>198</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>200</td>
<td>0.292</td>
<td>0.13</td>
<td>0.097</td>
<td>0.183</td>
<td>0.296</td>
<td>0.199</td>
<td>0.235</td>
<td>0.275</td>
</tr>
<tr>
<td>202</td>
<td>0.009</td>
<td>0.13</td>
<td>0.017</td>
<td>0.134</td>
<td>0.032</td>
<td>0.084</td>
<td>0.105</td>
<td>0.045</td>
</tr>
<tr>
<td>204</td>
<td>0.066</td>
<td>0.506</td>
<td>0.273</td>
<td>0.052</td>
<td>0.101</td>
<td>0.376</td>
<td>0.155</td>
<td>0.169</td>
</tr>
<tr>
<td>206</td>
<td>0.019</td>
<td>0.123</td>
<td>0.011</td>
<td>0.078</td>
<td>0.011</td>
<td>0.156</td>
<td>0.04</td>
<td>0.022</td>
</tr>
<tr>
<td>208</td>
<td>0.16</td>
<td>0.013</td>
<td>0.045</td>
<td>0.02</td>
<td>0.077</td>
<td>0.032</td>
<td>0.07</td>
<td>0.051</td>
</tr>
<tr>
<td>210</td>
<td>0.434</td>
<td>0.006</td>
<td>0.477</td>
<td>0.363</td>
<td>0.433</td>
<td>0.068</td>
<td>0.335</td>
<td>0.399</td>
</tr>
<tr>
<td>212</td>
<td>0.009</td>
<td>0</td>
<td>0.01</td>
<td>0.031</td>
<td>0.002</td>
<td>0.025</td>
<td>0.006</td>
<td>0</td>
</tr>
<tr>
<td>214</td>
<td>0.009</td>
<td>0</td>
<td>0.01</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>218</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>59</td>
<td>77</td>
<td>88</td>
<td>153**</td>
<td>647</td>
<td>223</td>
<td>99**</td>
<td>88</td>
</tr>
<tr>
<td>------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>----</td>
</tr>
<tr>
<td>251</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>253</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>255</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>257</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>259</td>
<td>0.008</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0.002</td>
<td>0.01</td>
<td>0.006</td>
</tr>
<tr>
<td>261</td>
<td>0.347</td>
<td>0.071</td>
<td>0.199</td>
<td>0.526</td>
<td>0.331</td>
<td>0.083</td>
<td>0.242</td>
<td>0.301</td>
</tr>
<tr>
<td>263</td>
<td>0.127</td>
<td>0.045</td>
<td>0.432</td>
<td>0.052</td>
<td>0.146</td>
<td>0.052</td>
<td>0.172</td>
<td>0.267</td>
</tr>
<tr>
<td>265</td>
<td>0.051</td>
<td>0.026</td>
<td>0.063</td>
<td>0.007</td>
<td>0.076</td>
<td>0.056</td>
<td>0.04</td>
<td>0.051</td>
</tr>
<tr>
<td>267</td>
<td>0.153</td>
<td>0.097</td>
<td>0.085</td>
<td>0.088</td>
<td>0.138</td>
<td>0.209</td>
<td>0.146</td>
<td>0.097</td>
</tr>
<tr>
<td>269</td>
<td>0.017</td>
<td>0.071</td>
<td>0.006</td>
<td>0.029</td>
<td>0.044</td>
<td>0.074</td>
<td>0.025</td>
<td>0.006</td>
</tr>
<tr>
<td>271</td>
<td>0.025</td>
<td>0.221</td>
<td>0</td>
<td>0.01</td>
<td>0.031</td>
<td>0.17</td>
<td>0.081</td>
<td>0.023</td>
</tr>
<tr>
<td>273</td>
<td>0.025</td>
<td>0.117</td>
<td>0</td>
<td>0.026</td>
<td>0.037</td>
<td>0.101</td>
<td>0.04</td>
<td>0.028</td>
</tr>
<tr>
<td>275</td>
<td>0.059</td>
<td>0.149</td>
<td>0</td>
<td>0.052</td>
<td>0.053</td>
<td>0.121</td>
<td>0.071</td>
<td>0.045</td>
</tr>
<tr>
<td>277</td>
<td>0.076</td>
<td>0.091</td>
<td>0.011</td>
<td>0.033</td>
<td>0.043</td>
<td>0.034</td>
<td>0.051</td>
<td>0.04</td>
</tr>
<tr>
<td>279</td>
<td>0.093</td>
<td>0.045</td>
<td>0.006</td>
<td>0.052</td>
<td>0.053</td>
<td>0.045</td>
<td>0.051</td>
<td>0.051</td>
</tr>
<tr>
<td>281</td>
<td>0.017</td>
<td>0.019</td>
<td>0</td>
<td>0.056</td>
<td>0.022</td>
<td>0.016</td>
<td>0.015</td>
<td>0</td>
</tr>
<tr>
<td>283</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.007</td>
<td>0.004</td>
<td>0</td>
<td>0.011</td>
</tr>
<tr>
<td>285</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0.023</td>
<td>0.009</td>
<td>0.009</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>287</td>
<td>0</td>
<td>0.006</td>
<td>0.193</td>
<td>0.016</td>
<td>0.002</td>
<td>0.009</td>
<td>0.045</td>
<td>0.063</td>
</tr>
<tr>
<td>289</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>293</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>295</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>297</td>
<td>0</td>
<td>0.019</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>58</td>
<td>78</td>
<td>87</td>
<td>152</td>
<td>623</td>
<td>215**</td>
<td>99</td>
<td>87</td>
</tr>
<tr>
<td>----------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-------</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>D5S410</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>323</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>325</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>329</td>
<td>0</td>
<td>0.019</td>
<td>0</td>
<td>0.003</td>
<td>0.002</td>
<td>0.016</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>331</td>
<td>0.121</td>
<td>0.019</td>
<td>0.017</td>
<td>0.132</td>
<td>0.1</td>
<td>0.047</td>
<td>0.066</td>
<td>0.08</td>
</tr>
<tr>
<td>333</td>
<td>0.293</td>
<td>0.154</td>
<td>0.816</td>
<td>0.592</td>
<td>0.269</td>
<td>0.228</td>
<td>0.288</td>
<td>0.379</td>
</tr>
<tr>
<td>335</td>
<td>0.138</td>
<td>0.628</td>
<td>0.103</td>
<td>0.066</td>
<td>0.082</td>
<td>0.481</td>
<td>0.237</td>
<td>0.132</td>
</tr>
<tr>
<td>337</td>
<td>0.25</td>
<td>0.135</td>
<td>0.052</td>
<td>0.184</td>
<td>0.372</td>
<td>0.179</td>
<td>0.242</td>
<td>0.236</td>
</tr>
<tr>
<td>339</td>
<td>0.034</td>
<td>0.026</td>
<td>0</td>
<td>0.01</td>
<td>0.021</td>
<td>0.021</td>
<td>0.035</td>
<td>0.023</td>
</tr>
<tr>
<td>341</td>
<td>0.121</td>
<td>0.006</td>
<td>0</td>
<td>0.003</td>
<td>0.011</td>
<td>0.007</td>
<td>0.015</td>
<td>0.006</td>
</tr>
<tr>
<td>343</td>
<td>0.095</td>
<td>0.006</td>
<td>0.011</td>
<td>0.007</td>
<td>0.099</td>
<td>0.009</td>
<td>0.071</td>
<td>0.098</td>
</tr>
<tr>
<td>345</td>
<td>0.034</td>
<td>0</td>
<td>0</td>
<td>0.016</td>
<td>0.002</td>
<td>0.025</td>
<td>0.011</td>
<td>0.01</td>
</tr>
<tr>
<td>347</td>
<td>0.034</td>
<td>0.006</td>
<td>0</td>
<td>0.016</td>
<td>0.005</td>
<td>0.015</td>
<td>0.029</td>
<td>0.029</td>
</tr>
<tr>
<td>349</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>351</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.008</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D7S246</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>116</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.011</td>
<td>0.045</td>
<td>0.009</td>
<td>0.027</td>
<td>0.022</td>
</tr>
<tr>
<td>118</td>
<td>0.044</td>
<td>0</td>
<td>0</td>
<td>0.008</td>
<td>0.064</td>
<td>0.018</td>
<td>0.043</td>
<td>0.021</td>
</tr>
<tr>
<td>120</td>
<td>0.044</td>
<td>0.426</td>
<td>0.4</td>
<td>0.934</td>
<td>0.561</td>
<td>0.429</td>
<td>0.459</td>
<td>0.522</td>
</tr>
<tr>
<td>122</td>
<td>0.044</td>
<td>0.016</td>
<td>0.007</td>
<td>0.022</td>
<td>0.057</td>
<td>0.021</td>
<td>0.02</td>
<td>0.022</td>
</tr>
<tr>
<td>124</td>
<td>0</td>
<td>0.082</td>
<td>0</td>
<td>0.004</td>
<td>0.009</td>
<td>0.041</td>
<td>0.014</td>
<td>0</td>
</tr>
<tr>
<td>126</td>
<td>0.022</td>
<td>0.115</td>
<td>0.007</td>
<td>0</td>
<td>0.012</td>
<td>0.1</td>
<td>0.041</td>
<td>0.015</td>
</tr>
<tr>
<td>128</td>
<td>0.111</td>
<td>0.197</td>
<td>0.436</td>
<td>0.011</td>
<td>0.106</td>
<td>0.215</td>
<td>0.155</td>
<td>0.235</td>
</tr>
<tr>
<td>130</td>
<td>0.044</td>
<td>0.139</td>
<td>0.014</td>
<td>0</td>
<td>0.004</td>
<td>0.082</td>
<td>0.02</td>
<td>0.015</td>
</tr>
<tr>
<td>132</td>
<td>0.011</td>
<td>0.016</td>
<td>0.007</td>
<td>0</td>
<td>0.009</td>
<td>0.038</td>
<td>0.034</td>
<td>0.015</td>
</tr>
<tr>
<td>134</td>
<td>0.011</td>
<td>0</td>
<td>0.05</td>
<td>0</td>
<td>0.051</td>
<td>0.015</td>
<td>0.081</td>
<td>0.059</td>
</tr>
<tr>
<td>136</td>
<td>0.067</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.027</td>
<td>0.006</td>
<td>0.02</td>
<td>0.029</td>
</tr>
<tr>
<td>138</td>
<td>0</td>
<td>0</td>
<td>0.014</td>
<td>0</td>
<td>0.075</td>
<td>0.024</td>
<td>0.068</td>
<td>0.015</td>
</tr>
<tr>
<td>140</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>144</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>65**</td>
<td>85</td>
<td>145**</td>
<td>634</td>
<td>222</td>
<td>100</td>
<td>88</td>
</tr>
<tr>
<td>------</td>
<td>----</td>
<td>------</td>
<td>----</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>D16S30</td>
<td>151</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>153</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>155</td>
<td>0</td>
<td>0.077</td>
<td>0</td>
<td>0.005</td>
<td>0.005</td>
<td>0.006</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>157</td>
<td>0.017</td>
<td>0.123</td>
<td>0</td>
<td>0.007</td>
<td>0.003</td>
<td>0.009</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>159</td>
<td>0.05</td>
<td>0.015</td>
<td>0.012</td>
<td>0.014</td>
<td>0.016</td>
<td>0.034</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>161</td>
<td>0.167</td>
<td>0.108</td>
<td>0.006</td>
<td>0.09</td>
<td>0.188</td>
<td>0.077</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>163</td>
<td>0.167</td>
<td>0.431</td>
<td>0.6</td>
<td>0.097</td>
<td>0.267</td>
<td>0.428</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>165</td>
<td>0.383</td>
<td>0.154</td>
<td>0.135</td>
<td>0.514</td>
<td>0.342</td>
<td>0.146</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>166</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>167</td>
<td>0.15</td>
<td>0.069</td>
<td>0.206</td>
<td>0.238</td>
<td>0.106</td>
<td>0.133</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>169</td>
<td>0.05</td>
<td>0.008</td>
<td>0</td>
<td>0.028</td>
<td>0.063</td>
<td>0.027</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>171</td>
<td>0.017</td>
<td>0.015</td>
<td>0</td>
<td>0.01</td>
<td>0.009</td>
<td>0.011</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>173</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>175</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>56</th>
<th>77**</th>
<th>85</th>
<th>152</th>
<th>626</th>
<th>221</th>
<th>101</th>
<th>83</th>
</tr>
</thead>
<tbody>
<tr>
<td>D10S17</td>
<td>183</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>185</td>
<td>0</td>
<td>0.162</td>
<td>0.012</td>
<td>0.092</td>
<td>0.002</td>
<td>0.131</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>187</td>
<td>0.098</td>
<td>0.123</td>
<td>0.141</td>
<td>0.257</td>
<td>0.149</td>
<td>0.084</td>
<td>0.193</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>0.232</td>
<td>0.279</td>
<td>0.006</td>
<td>0.082</td>
<td>0.149</td>
<td>0.217</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>191</td>
<td>0.018</td>
<td>0.084</td>
<td>0.018</td>
<td>0.105</td>
<td>0.011</td>
<td>0.054</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>193</td>
<td>0.027</td>
<td>0.182</td>
<td>0.488</td>
<td>0.141</td>
<td>0.075</td>
<td>0.174</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td>195</td>
<td>0.045</td>
<td>0.058</td>
<td>0.006</td>
<td>0.069</td>
<td>0.04</td>
<td>0.109</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>197</td>
<td>0.527</td>
<td>0.091</td>
<td>0.312</td>
<td>0.247</td>
<td>0.526</td>
<td>0.163</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td>199</td>
<td>0.054</td>
<td>0.019</td>
<td>0.018</td>
<td>0.007</td>
<td>0.046</td>
<td>0.068</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>203</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>59</th>
<th>77</th>
<th>88</th>
<th>150</th>
<th>630</th>
<th>215</th>
<th>97</th>
<th>88</th>
</tr>
</thead>
<tbody>
<tr>
<td>D15S10</td>
<td>208</td>
<td>0.034</td>
<td>0</td>
<td>0</td>
<td>0.017</td>
<td>0.026</td>
<td>0</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>210</td>
<td>0.398</td>
<td>0.104</td>
<td>0.244</td>
<td>0.123</td>
<td>0.361</td>
<td>0.133</td>
<td>0.247</td>
</tr>
<tr>
<td></td>
<td>212</td>
<td>0.144</td>
<td>0.078</td>
<td>0.193</td>
<td>0.343</td>
<td>0.101</td>
<td>0.056</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>214</td>
<td>0.034</td>
<td>0.013</td>
<td>0.017</td>
<td>0.013</td>
<td>0.01</td>
<td>0.037</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>216</td>
<td>0</td>
<td>0.065</td>
<td>0</td>
<td>0.01</td>
<td>0.001</td>
<td>0.065</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>218</td>
<td>0</td>
<td>0.299</td>
<td>0</td>
<td>0.003</td>
<td>0.01</td>
<td>0.27</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>220</td>
<td>0.025</td>
<td>0.136</td>
<td>0.006</td>
<td>0.107</td>
<td>0.02</td>
<td>0.079</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>222</td>
<td>0.22</td>
<td>0.065</td>
<td>0.04</td>
<td>0.147</td>
<td>0.274</td>
<td>0.091</td>
<td>0.237</td>
</tr>
<tr>
<td></td>
<td>224</td>
<td>0.119</td>
<td>0.143</td>
<td>0.136</td>
<td>0.177</td>
<td>0.135</td>
<td>0.172</td>
<td>0.165</td>
</tr>
<tr>
<td></td>
<td>226</td>
<td>0.008</td>
<td>0.045</td>
<td>0.33</td>
<td>0.04</td>
<td>0.032</td>
<td>0.044</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>228</td>
<td>0.008</td>
<td>0.032</td>
<td>0.011</td>
<td>0.007</td>
<td>0.02</td>
<td>0.03</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>230</td>
<td>0.008</td>
<td>0.006</td>
<td>0.023</td>
<td>0</td>
<td>0.011</td>
<td>0.009</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>232</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>234</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>236</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.009</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>59</td>
<td>77</td>
<td>88</td>
<td>154</td>
<td>642</td>
<td>219</td>
<td>100</td>
<td>89</td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>D6S161</td>
<td>267</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>269</td>
<td>0</td>
<td>0.039</td>
<td>0</td>
<td>0</td>
<td>0.046</td>
<td>0.005</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>271</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>273</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>275</td>
<td>0.025</td>
<td>0.019</td>
<td>0.097</td>
<td>0.003</td>
<td>0.051</td>
<td>0.032</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>277</td>
<td>0.178</td>
<td>0.214</td>
<td>0.023</td>
<td>0.357</td>
<td>0.199</td>
<td>0.164</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>279</td>
<td>0.093</td>
<td>0.279</td>
<td>0.188</td>
<td>0.198</td>
<td>0.104</td>
<td>0.263</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>281</td>
<td>0.229</td>
<td>0.091</td>
<td>0.091</td>
<td>0.036</td>
<td>0.252</td>
<td>0.094</td>
<td>0.275</td>
</tr>
<tr>
<td></td>
<td>283</td>
<td>0.153</td>
<td>0.045</td>
<td>0.034</td>
<td>0.159</td>
<td>0.123</td>
<td>0.066</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>284</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>285</td>
<td>0.254</td>
<td>0.091</td>
<td>0.307</td>
<td>0.224</td>
<td>0.203</td>
<td>0.153</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>286</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>287</td>
<td>0.068</td>
<td>0.026</td>
<td>0.25</td>
<td>0.016</td>
<td>0.048</td>
<td>0.068</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>289</td>
<td>0</td>
<td>0.026</td>
<td>0.011</td>
<td>0.006</td>
<td>0.004</td>
<td>0.039</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>291</td>
<td>0</td>
<td>0.084</td>
<td>0</td>
<td>0</td>
<td>0.008</td>
<td>0.025</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>293</td>
<td>0</td>
<td>0.045</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.039</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>295</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0.004</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>297</td>
<td>0</td>
<td>0.013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>58</th>
<th>78</th>
<th>86</th>
<th>147</th>
<th>624</th>
<th>210</th>
<th>102</th>
<th>87</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1S262</td>
<td>337</td>
<td>0.009</td>
<td>0.006</td>
<td>0.012</td>
<td>0.197</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>339</td>
<td>0.319</td>
<td>0.083</td>
<td>0</td>
<td>0.01</td>
<td>0.286</td>
<td>0.138</td>
<td>0.343</td>
</tr>
<tr>
<td></td>
<td>341</td>
<td>0.026</td>
<td>0.16</td>
<td>0.134</td>
<td>0.014</td>
<td>0.024</td>
<td>0.1</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>343</td>
<td>0.483</td>
<td>0.096</td>
<td>0.57</td>
<td>0.548</td>
<td>0.479</td>
<td>0.193</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>345</td>
<td>0.043</td>
<td>0.244</td>
<td>0.023</td>
<td>0.007</td>
<td>0.073</td>
<td>0.236</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>347</td>
<td>0.121</td>
<td>0.327</td>
<td>0.238</td>
<td>0.218</td>
<td>0.128</td>
<td>0.281</td>
<td>0.152</td>
</tr>
<tr>
<td></td>
<td>349</td>
<td>0</td>
<td>0.077</td>
<td>0.023</td>
<td>0.007</td>
<td>0.002</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>351</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>353</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D12S35</td>
<td>57</td>
<td>69</td>
<td>87</td>
<td>150</td>
<td>627</td>
<td>214</td>
<td>101</td>
<td>88</td>
</tr>
<tr>
<td>--------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>145</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>147</td>
<td>0</td>
<td>0.029</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0.016</td>
<td>0.01</td>
<td>0.006</td>
</tr>
<tr>
<td>149</td>
<td>0</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>151</td>
<td>0.009</td>
<td>0</td>
<td>0.023</td>
<td>0.013</td>
<td>0.049</td>
<td>0.009</td>
<td>0.035</td>
<td>0.04</td>
</tr>
<tr>
<td>153</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
</tr>
<tr>
<td>155</td>
<td>0</td>
<td>0.007</td>
<td>0.011</td>
<td>0.003</td>
<td>0.004</td>
<td>0.002</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>157</td>
<td>0.053</td>
<td>0.058</td>
<td>0.287</td>
<td>0.18</td>
<td>0.088</td>
<td>0.082</td>
<td>0.054</td>
<td>0.125</td>
</tr>
<tr>
<td>159</td>
<td>0.061</td>
<td>0.072</td>
<td>0.362</td>
<td>0.017</td>
<td>0.051</td>
<td>0.091</td>
<td>0.119</td>
<td>0.136</td>
</tr>
<tr>
<td>161</td>
<td>0.114</td>
<td>0.246</td>
<td>0.023</td>
<td>0.067</td>
<td>0.126</td>
<td>0.301</td>
<td>0.104</td>
<td>0.091</td>
</tr>
<tr>
<td>163</td>
<td>0.518</td>
<td>0.413</td>
<td>0.121</td>
<td>0.36</td>
<td>0.477</td>
<td>0.364</td>
<td>0.465</td>
<td>0.364</td>
</tr>
<tr>
<td>165</td>
<td>0.228</td>
<td>0.058</td>
<td>0.08</td>
<td>0.26</td>
<td>0.186</td>
<td>0.079</td>
<td>0.188</td>
<td>0.159</td>
</tr>
<tr>
<td>167</td>
<td>0.009</td>
<td>0.08</td>
<td>0.069</td>
<td>0.057</td>
<td>0.014</td>
<td>0.042</td>
<td>0.005</td>
<td>0.051</td>
</tr>
<tr>
<td>169</td>
<td>0.009</td>
<td>0.022</td>
<td>0.017</td>
<td>0.03</td>
<td>0.005</td>
<td>0.012</td>
<td>0.015</td>
<td>0.017</td>
</tr>
<tr>
<td>171</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>173</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>175</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FY</th>
<th>56</th>
<th>73</th>
<th>80</th>
<th>152</th>
<th>568**</th>
<th>209</th>
<th>65</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.982</td>
<td>0.014</td>
<td>0.994</td>
<td>1</td>
<td>0.993</td>
<td>0.17</td>
<td>0.762</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>0.018</td>
<td>0.986</td>
<td>0.006</td>
<td>0</td>
<td>0.007</td>
<td>0.83</td>
<td>0.238</td>
<td>NA</td>
</tr>
</tbody>
</table>

First row shows the numbers of individuals genotyped for the allele. First column shows the alleles as no. of microsatellite repeats. EU: Spanish, WAF: Mende, IA: Nahua, AS: Thai, EACT: European Americans from CT, AACT: African Americans from CT, LCT: Latinos from CT, LCA: Latinos from California

\* na” Marker was not typed in and no genotypes were available
\* Marker not in HWE: Homozygote deficiency
\** Marker not in HWE: Heterozygote deficiency
Appendix B

Perl script for generating individuals with known admixture proportions

Sim_sample.pl

Program developed by Mark Shriver and Indrani Halder. Script written by Vibhor Sonpar

#Read in the format file
open(FILE,"format.txt");
$k=0;
while($data=<FILE>)
{
    print $data;
    chomp($data);
    $data =~ /Marker1=(\w*)/;
    $mlk[$k]=$1;
    $data=<FILE>
    $data=<FILE>
    chomp($data);
    $c=0;
    while($data ne "END")
    {
        $dat[$k][$c]=$data;
        $c++;
        $data=<FILE>
        chomp($data);
    }
    $cnt[$k]=$c;
    $k++;
    $data=<FILE>
    $data=<FILE>
}
close(FILE);

#inputfile
open(FILE,"inputfile.txt");
$indv=<FILE>
chomp($indv);
$mark=<FILE>
chomp($mark);
$gps=<FILE>
chomp($gps);
for($i=0;$i<$gps;$i++)
{
    $data=<FILE>
    chomp($data)
    $id[$i]=$data;
}
close(FILE);
open(OFILE,">output.txt");
for($i=0;$i<$mark;$i++)
{
    print OFILE $mlk[$i],"t";
}
print OFILE "n"

for($q=0;$q<$indv;$q++)
{
    print $q,"n";
    print OFILE "Indv","q,"t";
}\loop for markers
for($i=0;$i<$mark;$i++)
{
    \loop for ancestors
    for($j=0;$j<$gps;$j++)
    {
        \loop to get freq data
        for($m=0;$m<$cnt[$i];$m++)
        {
            @da=split(/\t/,$dat[$i][$m]);
            $ma[$m]=$da[0];
            $va[$m]=$da[$id[$j]];
        }
        \loop to get alleles
        for($m=0;$m<2;$m++)
        {
            $cnt=-1;
            $val=0;
            $a=rand;
            # TODO: ensure that $val add up to 1 (allele frequencies are
            normalized)
            while($a == 0)
            {
                $a=rand;
            }
            while($a > $val)
            {
                $cnt++;
                $val+=$va[$cnt];
            }
            $all[$j][$m]=$ma[$cnt];
        }
        print OFILE $all[$j];
    }
    \loop to go over generations
    $temp=$gps;
    while($temp > 1)
    {
        print OFILE "n";
        $ct=0;
        for($h=0;$h<$temp;$h=$h+2)
        {
            $all[$ct][0]=&reduce($all[$h][0],$all[$h][1]);
$all[$ct][1]=&reduce($all[$h+1][0],$all[$h+1][1]);
    #print OFILE $all[$ct];
    $ct++;
}
    $temp=$ct;
}
    #print OFILE "\n";
    print OFILE $all[0][0],"\t",$all[0][1],"\t";
}
print OFILE "\n";
}

close(OFILE);

# reducing within generations
sub reduce {
    my($v1,$v2);
    ($v1,$v2) = @_;  
    $a1=rand;
    if($a1 > 0.5) {
        $c1=$v1;
    } else {
        $c1=$v2;
    }
    return $c1;
}
Appendix C

Perl script for generating individuals with known admixture proportions and tracking alleles through generations

Sim_sample_tag.pl

Program developed by Mark Shriver and Indrani Halder. Script written by Vibhor Sonpar, modified by Xianyun Mao and Indrani Halder.

Program requires user to specify one file containing allele frequencies of ancestral populations (the “format.txt” file). A second file that has to be specified is ‘inputfile.txt’ which contains information on the number of individuals to be generated, number of generations the program should be used, number of markers used for the simulation and the order of individuals from the ancestral populations to establish the mating order in the first generation.

```perl
# Read in the format file
open(FILE, "format.txt");
$k=0;
while($data=<FILE>)
{
    chomp($data);
    @array=split("\t", $data);
    @markerid=split("=",$array[0]);
    $mlk[$k]=$markerid[1];
    $data=<FILE>;
    $data=<FILE>;
    chomp($data);
    $c=0;
    while($data!~/END/) 
    {
        $dat[$k][$c]=$data;
        $c++;
        $data=<FILE>;
    }
    chomp($data);
    $cnt[$k]=$c;
    $k++;
```
$data=<FILE>;
$data=<FILE>;
}
close(FILE);

#specifying the inputfile which contains all parameters for the
#simulation
open(FILE,"inputfile.txt");

$indv=<FILE>;
chomp($indv);
$mark=<FILE>;
chomp($mark);
$gps=<FILE>;
chomp($gps);
for($i=0;$i<$gps;$i++)
{
    $data=<FILE>;
    chomp($data);
    $id[$i]=$data;
}
close(FILE);

#introducing tags for all alleles and attaching tags to alleles
open(OFILE2,">output_tagged.txt");
open(OFILE1,">output.txt");
for($i=0;$i<$mark;$i++)
{
    print OFILE1 $mlk[$i],"\t";
    print OFILE2 $mlk[$i],"\t";
}
print OFILE1 "\n";
print OFILE2 "\n";
for($q=0;$q<$indv;$q++)
{
    print $q,"\n";
    print OFILE1 "Indv","$q","\t";
    print OFILE2 "Indv","$q","\t";

#loop for markers
for($i=0;$i<$mark;$i++)
{
    #loop for ancestors
    for($j=0;$j<$gps;$j++)
    {
        #loop to get freq data
        for($m=0;$m<$cnt[$i];$m++)
        {
            @da=split(/\t/,$dat[$i][$m]);
            $ma[$m]=@da[0];
            $va[$m]=@da[$id[$j]];
        }
    }
}
#check the allele frequencies
$sum=0;
for($m=0;$m<$cnt[$i];$m++)
{
    $sum=$sum+$va[$m];
}
if($sum==1) {
} else {
    for($m=0;$m<$cnt[$i];$m++)
    {
        $va[$m]=1/$cnt[$i];
    }
}
# loop to get alleles
for($m=0;$m<2;$m++)
{
    $cnt=-1;
    $val=0;
    $a=rand;
    # TODO: ensure that $val add up to 1 (allele frequencies are normalized)
    while($a == 0)
    {
        $a=rand;
    }
    while($a > $val)
    {
        $cnt++;
        $val+=$va[$cnt];
    }
    $allt[$j][$m]=&ma[$cnt]."p$id[$j]";
}
#print OFILE $all[$j];
}
# loop to go over generations
$temp=$gps;
while($temp > 1)
{
    #print OFILE "\n";
    $ct=0;
    for($h=0;$h<$temp;$h=$h+2)
    {
        $allt[$ct][0]=&reduce($allt[$h][0],$allt[$h][1]);
        $allt[$ct][1]=&reduce($allt[$h+1][0],$allt[$h+1][1]);
        #print OFILE $all[$ct];
        $ct++;}
    $temp=$ct;
}
#print OFILE "\n";
@array1=split("p",$allt[0][0]);
@array2=split("p",$allt[0][1]);
print OFILE1 $array1[0],"\t",$array2[0],"\t";
print OFILE2 $allt[0][0],"\t",$allt[0][1],"\t";

}
print OFILE1 "\n";
print OFILE2 "\n";
}
close(OFILE1);
close(OFILE2);

# summarize segregation ancestry proportions;

open(ifile,"output_tagged.txt");
open(ofile,">absolute_admix.txt");
cline=<ifile>;
print $line;
$pops=0;
$pop[$pops]=$id[0];
# count the number of source pops;
for($i=1;$i<$gps;$i++) {
$found=0;
for($j=0;$j<=$($i-1);$j++) {
  if($id[$i]==$id[$j]) {
    $found=1;
    last;
  }
}
if($found==1) { } else { 
  $pops++;
  $pop[$pops]=$id[$i];
  }
}
for($i=0;$i<=$pops;$i++) {
  print ofile "\tpop$pop[$i];"
}
print ofile "\n";

while(eof(ifile)==0) {
  $line=<ifile>;
  chomp($line);
  @array=split("\t",$line);
  print ofile "$array[0]";
  for($i=0;$i<=$pops;$i++) {
    $count[$i]=0;
    $popcall[$i]="p".$pop[$i];
  }
  for($i=1;$i<=$#array;$i++) {
    for($j=0;$j<=$pops;$j++) {
      if($array[$i]!~/^$popcall[$j]$/) {
for($i=0;$i<=$pops;$i++) {
    $freq=$count[$i]/(2*$mark);
    $freq=substr($freq,0,6);
    print ofile "\t$freq"
    }

    print ofile "\n"
}
close(ifile);
close(ofile);

# reducing within generations
sub reduce {
    my($v1,$v2);
    ($v1,$v2) = @_;  # $x = @_;
    $a1=rand;
    if($a1 > 0.5) {  # $a1 > 0.5 { $c1=$v1; }
        $c1=$v1;
    } else {          # $a1 > 0.5 { $c1=$v2; }
        $c1=$v2;
    }
    return $c1;
}
Appendix D

Analyzing individual BioGeographical ancestry and Admixture structure in human populations using 10K markers

The analyses described in this appendix have been published as part of Shriver et al. (2005), which describes continuous and clustered genetic variation observed among different populations.

Understanding genetic variation in populations and how this variation contributes to differences in disease risk among populations is an important area of research. We have typed a panel of 11,078 SNP markers in three resident US populations, (African Americans, Puerto Ricans and European Americans) in an effort to analyze patterns of genetic variability. The first two are clearly admixed populations with varying levels of BioGeographical ancestry (BGA), while the European Americans are more homogeneous. Putative ancestral populations were chosen from among a panel of twelve world populations typed for the same markers. Using a maximum likelihood (ML) based method and a separate Bayesian method we show that individual BGA estimates vary within all groups. Assuming a trihybrid model of admixture between Europeans, West Africans and Native Americans we observed large variations in individual BGA levels in both the African-American and Puerto Rican samples, while the European Americans are more tightly clustered.

To test for stratification within the population samples we have used the Individual Ancestry Correlation (IAC) test, where markers are split into two non-overlapping subsets and ancestry is estimated separately with both sets. Correlation between estimates obtained with the different marker sets indicates presence of genetic structure. While both the African-American and the Puerto Rican samples show admixture stratification as expected, we also detected evidence for stratification within the European-American sample.
Materials and Methods:

The study populations included 41 European Americans (EA), 42 African Americans (AA) from Coriel repository and 20 Puerto Ricans (PR) from New York City. Ancestral allele frequencies were estimated from 22 Mende from Sierra Leone, 20 Spanish from Valencia, Spain, 20 Nahua from Peru and 20 Quechua from Mexico.

All individuals were genotyped for 11,555 markers using the Affymetrix 10K WGA chip (Kennedy et al. 2003). Maximum likelihood (ML) method as described in (Hanis et al. 1986) was implemented through Perl scripts. Statistical analysis of data was carried out using standard statistical packages. STRUCTURE 2.0 (Pritchard et al. 2000; Falush et al. 2003) was used separately to estimate ancestry, assuming an admixture model, uncorrelated allele frequencies, independent alpha parameter, with 30,000 burn in and 70,000 repetitions for collecting parameter estimates.

Results:

Table 1: IAE : ML

<table>
<thead>
<tr>
<th>Population</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>0.06 ± 0.026</td>
<td>0.21 ± 0.081</td>
<td>0.73 ± 0.093</td>
</tr>
<tr>
<td>European American</td>
<td>0.088 ± 0.014</td>
<td>0.831 ± 0.019</td>
<td>0.081 ± 0.021</td>
</tr>
<tr>
<td>Puerto Rican</td>
<td>0.178 ± 0.393</td>
<td>0.586 ± 0.148</td>
<td>0.236 ± 0.171</td>
</tr>
</tbody>
</table>

Mean ± Standard Deviation of ancestry estimates in different populations using ML
Figure 1: Proportional ancestry in three different population samples using ML. Filled circles: European Americans, blank circles: African Americans; filled triangles: Puerto Ricans.
Proportional Ancestry in European Americans

Figure 2a: Plot showing distribution of Individual Ancestry in European Americans using ML

Proportional Ancestry in African Americans

Figure 2b: Plot showing distribution of Individual Ancestry in African Americans using ML
The ML estimates of proportional ancestry show a greater tendency for the European-American subjects to cluster together as compared to the other two population samples. The African Americans (AA) and Puerto Ricans (PR) both show greater variability in individual ancestry levels with most non-African ancestry in the AA being from Europe. The PR sample shows some individuals with more Indigenous American ancestry as well as substantial West African ancestry.

Table 2: IAE: STRUCTURE:

<table>
<thead>
<tr>
<th>Population</th>
<th>% Indigenous American</th>
<th>% European</th>
<th>% West African</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>0.01 ± 0.02</td>
<td>0.19 ± 0.09</td>
<td>0.79 ± 0.12</td>
</tr>
<tr>
<td>European American</td>
<td>0.22 ± 0.02</td>
<td>0.76 ± 0.02</td>
<td>0.01 ± 0.01</td>
</tr>
<tr>
<td>Puerto Rican</td>
<td>0.16 ± 0.04</td>
<td>0.63 ± 0.16</td>
<td>0.22 ± 0.19</td>
</tr>
</tbody>
</table>

Mean ± Standard Deviation of ancestry estimates in different populations using ML

STRUCTURE was run assuming an admixture model with uncorrelated allele frequencies, individual alpha parameters, with a burn in period of 30,000 and 70,000 repetitions. Individuals representing ancestral populations were included in all analyses.
Figure 3: Proportional ancestry observed in European Americans (Filled circles), African American (Grey circles) and Puerto Ricans (Filled triangles)
Distribution of proportional ancestry using STRUCTURE

Propotional ancestry in European Americans using STRUCTURE

Figure 4a: Plot showing distribution of Individual Ancestry in European Americans using ML.
Figure 4b: Plot showing distribution of Individual Ancestry in African Americans using ML
Figure 4c: Plot showing distribution of Individual Ancestry in European Americans using ML.

Table 3: Correlations in estimates obtained with ML and STRUCTURE

<table>
<thead>
<tr>
<th>Population</th>
<th>IA</th>
<th>EU</th>
<th>WAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>0.93</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>European American</td>
<td>0.8</td>
<td>0.67</td>
<td>0.99</td>
</tr>
<tr>
<td>Puerto Rican</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Shown is Spearman’s correlation coefficient. In all cases P<0.0001

Strong correlations were observed between estimates obtained with the two methods, especially in the African American and Puerto Rican sample. Correlations were lower in European Americans.

Tests for the presence of admixture structure are based on correlations in individual ancestry indices calculated from independent (unlinked) panels of markers. We tested for significant correlations using two types of individual indices, principal coordinates and BGA, calculated separately from the even and odd chromosomal SNPs.
In the absence of structure within a population, which is related to the axes of ancestry measured by these indices, no significant relationship between the two estimates is expected.

Table 4: Results of the Individual ancestry correlation test.

<table>
<thead>
<tr>
<th>Population</th>
<th>% Indigenous American</th>
<th>% European</th>
<th>% West African</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>0.95</td>
<td>0.9</td>
<td>0.64</td>
</tr>
<tr>
<td>European American</td>
<td>0.77</td>
<td>0.757</td>
<td>0.4</td>
</tr>
<tr>
<td>Puerto Rican</td>
<td>0.88</td>
<td>0.92</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Shown are Spearman’s correlation coefficient values. P values were all <0.001 in each case.

Table 5: Results of PC-correlation test.

<table>
<thead>
<tr>
<th>Population</th>
<th>First PC</th>
<th>Second PC</th>
<th>Third PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>0.79 (&lt;0.001)</td>
<td>0.49 (&lt;0.001)</td>
<td>0.125 (NS)</td>
</tr>
<tr>
<td>European American</td>
<td>0.49 (&lt;0.001)</td>
<td>0.021 (NS)</td>
<td>0.143 (NS)</td>
</tr>
<tr>
<td>Puerto Rican</td>
<td>0.62 (0.003)</td>
<td>0.76 (&lt;0.001)</td>
<td>0.417 (NS)</td>
</tr>
</tbody>
</table>

Spearman’s correlation coefficients are shown in parenthesis. NS: Non significant.

Both the analyses show high levels of correlation between estimates obtained with different marker sets. While such correlations are expected in the African American and Puerto Rican samples, which are admixed populations, correlations were also observed in the European American sample, indicating presence of admixture structure in a sample assumed to be more homogeneous.

We have used 20 Spanish individuals as ancestral Europeans. This could explain partly why we observe European Americans moving away from the European vertex on the triangle plot. It is known that British, Germans, Irish and Italians have contributed significantly to the European American gene pool and a more accurate estimate of proportional ancestry in European Americans would be obtained if we had used a larger sample of different European populations.
References:


Appendix E

Perl script for calculating Maximum Likelihood estimates of individual genomic ancestry

This program was developed by Mark Shriver, written by Vibhor Sonpar and modified by Indrani Halder.

#User Input

#Please replace the values for the first and second ancestral Populations
# 1 – Indigenous American
# 2 - European
# 3 – West African

$pop1 = 1;
$pop2 = 2;
$pop3 = 3;

#This program is written to find the maximum likelihood estimates

#Open the frequency data file
open(FILE,"reqdformat.txt");

#Read in the frequency data file into a two dim array based on the marker name
$k=0;
$l=0;
while($data=<FILE>)
{
    chomp($data);
    if($data =~ /Marker1=((\W*\w*)+)/)
    {
        $mk[$k]=$1;
    }
    $temp=<FILE>;
    $tdata=<FILE>;
    chomp($tdata);
    while(!($tdata =~ /END/))
    {
        $l++;
        @dat=splitlet/\t/,$tdata;
        $num[$k][$l]=$dat[0];
        $p1[$k][$l]=$dat[$pop1];
        $p2[$k][$l]=$dat[$pop2];
        $p3[$k][$l]=$dat[$pop3];
    }
}
$tdata=<FILE>;
chomp($tdata);
}
$k[$k]=$l;
$k++;
$l=0;
$temp=<FILE>;
$temp=<FILE>;
} close(FILE);

#Open the input data file
open(FILE,"inputtomlike.txt");

#Read in the data storing in arrays based on the individual id and marker name
$p=0;
$data=<FILE>;
chomp($data);
@markers=split(/\t/,$data);
$len= @markers;

while($data=<FILE>)
{
    chomp($data);
    @da=split(/\t/,$data);
    $tlen=@da;
    $name[$p]=$da[0];
    $u=1;
    while($u<=$tlen)
    {
        $par1[$p][((($u-1)/2)]=da[$u];
        $par2[$p][((($u-1)/2)]=da[$u+1];
        $u+=2;
    }
    $p++;
}
close(FILE);

for($j=0;$j<=$len;$j++)
{
    for($o=0;$o<=$k;$o++)
    {
        if($markers[$j] eq $mk[$o])
        {
            for($i=0;$i<$p;$i++)
            {
                $flg=0;
                #Loop for the counts
                for($b=0;$b<=$cnt[$o];$b++)
                {

```
# Check for missing data
if (($par1[$i][$j] eq "?") || ($par2[$i][$j] eq "?"))
{
    $t1[$i][$j]=0;
    $t2[$i][$j]=0;
    $t3[$i][$j]=0;
    $t4[$i][$j]=0;
    $t5[$i][$j]=0;
    $t6[$i][$j]=0;
}

# Populate variables with the freq values of the two populations
if ($par1[$i][$j] eq $num[$o][$b])
{
    $flg=1;
    $t1[$i][$j]=$p1[$o][$b];
    $t2[$i][$j]=$p2[$o][$b];
    $t3[$i][$j]=$p3[$o][$b];
    if ($t1[$i][$j] == 0)
    {
        $t1[$i][$j]=0.0001;
    }
    if ($t2[$i][$j] == 0)
    {
        $t2[$i][$j]=0.0001;
    }
    if ($t3[$i][$j] == 0)
    {
        $t3[$i][$j]=0.0001;
    }
}

if ($par2[$i][$j] eq $num[$o][$b])
{
    $flg=1;
    $t4[$i][$j]=$p1[$o][$b];
    $t5[$i][$j]=$p2[$o][$b];
    $t6[$i][$j]=$p3[$o][$b];
    if ($t4[$i][$j] == 0)
    {
        $t4[$i][$j]=0.0001;
    }
    if ($t5[$i][$j] == 0)
    {
        $t5[$i][$j]=0.0001;
    }
    if ($t6[$i][$j] == 0)
    {
        $t6[$i][$j]=0.0001;
    }
}

if ($flg)
$valid[$i]++;
}
}
break;

# Loop for individuals
for($i=0;$i<$p;$i++) {
    # Loop for m1
    for($m1=0;$m1<=101;$m1++) {
        for($m2=0;$m2<(101-$m1);$m2++) {
            print $i,"\t",$m1,"\t",$m2,"\n";
            $m3=100-$m1-$m2;
            $sum[$i][$m1][$m2]=0;
            $ma1=$m1/100;
            $ma2=$m2/100;
            $ma3=$m3/100;
            for($j=0;$j<=$len;$j++) {
                # check if the two alleles are equal if they are then add to sum
                # sqr(m1*t1+m2*t2) else 2*(m1*t1+m2*t2)*(m1*t3+m2*t4)
                if($par1[$i][$j] eq $par2[$i][$j]) {
                    $tval=($ma1*$t1[$i][$j]+$ma2*$t2[$i][$j]+$ma3*$t3[$i][$j])*($ma1*$t1[$i][$j]+$ma2*$t2[$i][$j]+$ma3*$t3[$i][$j]);
                    if(!$tval == 0) {
                        $sum[$i][$m1][$m2]+=log($tval)/log(10);
                    }
                } else {
                    $tval=2*($ma1*$t4[$i][$j]+$ma2*$t5[$i][$j]+$ma3*$t6[$i][$j])*($ma1*$t4[$i][$j]+$ma2*$t5[$i][$j]+$ma3*$t6[$i][$j]);
                    print $tval,"\n";
                    if(!$tval == 0) {
                        $sum[$i][$m1][$m2]+=log($tval)/log(10);
                    }
                }
            }
        }
    }
}
# Print the whole file
open(OFILE,">maxlike3.txt");

for($i=0;$i<$p;$i++)
{
    print OFILE $name[$i],"\n";
    $max[$i]=-100000;
    for($m1=0;$m1<=101;$m1++)
    {
        for($m2=0;$m2<(101-$m1);$m2++)
        {
            print OFILE $m1/100,"\t",$m2/100,"\t",(100-$m1-$m2)/100,"\t";$sum[$i][$m1][$m2],"\n";
            if($sum[$i][$m1][$m2] > $max[$i])
            {
                $max[$i]=$sum[$i][$m1][$m2];
                $x[$i]=$m1/100;
                $y[$i]=$m2/100;
                $z[$i]=(100-$m1-$m2)/100;
            }
        }
    }
    print OFILE "\n\n";
}
close(OFILE);

# Print the maximum values
open(OFILE,">maxvalues3.txt");

for($i=0;$i<$p;$i++)
{
    print OFILE $name[$i],"\t",$x[$i],"\t",$y[$i],"\t",$z[$i],"\t",$max[$i],"\t";$valid[$i],"\n";
}
close(OFILE);
VITA

Indrani Halder

EDUCATION:
2002-2005  The Pennsylvania State University, University Park, PA 16802
PhD in Genetics (Advisor: Dr. Mark Shriver)

2000-2002  The Pennsylvania State University, University Park, PA 16802
MS in Genetics (Advisor: Dr. Richard Ordway)

1996-1998  University of Calcutta, India
MSc in Zoology, Specialization in Genetics

1993-1996  University of Calcutta, India
BSc with Honors in Zoology

MEMBERSHIPS:
American Society for Human Genetics
International Genetic Epidemiology Society
International Society for Computational Biology

SELECTED PRESENTATIONS AND PUBLICATIONS:


