The Pennsylvania State University
The Graduate School
College of the Liberal Arts

POPULATION GENOMIC STRUCTURE OF EUROPE: INFORMATIVE MARKERS, METHODS AND PHENOTYPES

A Thesis in
Anthropology
by
Marc Philippe Bauchet

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2007

The thesis of Marc P. Bauchet was reviewed and approved* by:

Mark D. Shriver
Associate Professor of Biological anthropology
Thesis Advisor
Chair of Committee

Hiroshi Akashi
Associate Professor of Biology

Alan Walker
Evan Pugh Professor of Biological Anthropology and Biology

Kenneth M. Weiss
Evan Pugh Professor of Biological Anthropology and Genetics

Nina Jablonski
Professor and Head of the Department of Anthropology

*Signatures are on file at the Graduate School

ABSTRACT

Among continents Europe is remarkable for its relative small size, dearth of migration barriers and abundance of historical population movements. These features are compatible with the observation of small inter-population genetic distances and relative lack of population structure compared to other continents. However some phenotypic features such as lightness in hair, skin and eye pigmentation are specific to Europe and diverse throughout the continent. I present a survey of variation of these phenotypes among population samples collected in 2005 and 2006 in the US and Europe (by me and collaborators, with proper consent from participants), including some 1,000 European Americans and 720 Europeans, and focusing on samples of French, US Ashkenazi Jews, Irish, Italian, Polish and Portuguese descent (respectively 169, 39, 147, 140, 84 and 187 individuals). In order to study genetic variation in Europe a variety of population samples without phenotypes was also assembled. These samples from Europe and neighboring regions were obtained from various sources, including the CEPH-HGDP, the Coriell Institute, and many collaborators. The genetic ancestry and population structure of Europe were first investigated through a commercial genetic test, EuroDNA 1.0 by the company DNAPrint (Sarasota, FL). In parallel this allowed evaluating the usefulness and performance of this test to predict individual ancestry and possibly pigmentation phenotypes; both possibilities appeared rather limited, mainly because of the low informativeness and size of the marker panel used. I then report on genome-wide typing of 297 individuals for *ca.* 10,000 (10k) single nucleotide polymorphisms (SNPs)—using Affymetrix (Santa Clara, CA) 10k mapping array; the data reveal significant axes of population structure in Europeans of known and unknown ancestry, mainly differentiating Iberians, northern Europeans and southeastern Europeans, as well as Basque and Finnish individuals. A proper understanding of population genetic stratification—differences in individual ancestry within or among populations—is crucial in attempts to find genes for complex traits through association mapping. This section using the 10k array demonstrates the selection and application of EuroAIMs (European Ancestry Informative Markers) for ancestry estimation and correction of stratification, using validated Bayesian analysis (*structure* program) and Principal Coordinate Analysis (PCoA) from individual allele sharing distances. Based on the latter I present methods for detecting and measuring genetic population structure in individual population samples—one based on ANOVA, the other being an expansion of a split-half reliability method. The Coriell “Caucasian” and CEPH Utah sample panels, often used as proxies for European populations, are found to reflect different subsets of the continent’s ancestry.

Finally, using Illumina 317k mapping array data, both from 29 population sample pool and 180 individual genotypes (mainly from Ireland, Italy, Poland and Portugal), I confirm and expand the description of European population structure. A preliminary panel of 12k EuroAIMs was selected from the population pools, and applied to the 180 European individuals, which cluster apart based on country of origin (Ireland, Italy, Poland and Portugal). Individuals from other countries are also presented in this context. The availability of pigmentation phenotypes for these individuals will allow gene association studies and possibly admixture mapping. I finally discuss the potential of such data for biomedical research and understanding human genetic variation, as well as the danger in overstating commercial genetic test results to the public.

TABLE OF CONTENTS

LIST OF FIGURES vii
LIST OF TABLES x
ACKNOWLEDGEMENTS xi
Chapter 1 Introduction 1
European Population 2
Europe before anatomically modern *Homo sapiens* 3
The Neolithic revolution 4
Genetic Markers 5
Methods 6
Phenotypes 7
Stratification 7
Admixture v. clines 8
Chapter 2 European Genetic Ancestry and Pigmentation Phenotypes 9
Introduction 10
Material and Methods 13
Population Sampling 13
Eliminating Outliers and Relatives 15
The EuroDNA 1.0 Test 15
Genetic Distance 17
Principal Coordinate Analysis 17
PC axis significance 17
Bayesian analysis 19
Phenotypic measures 20
Data management 20
Statistical Analysis 21
Results 22
DNAPrint Ancestry Tests 22
Raw Genotypes Analyses with *structure* 23
Raw Genotypes Analyses with PCoA 26
Phenotypes 29
Hair Pigmentation 29
Skin Pigmentation 31
Eye Pigmentation 33
Phenotype-Ancestry Correlation 35
Hair Pigmentation 35
Skin Pigmentation 37
Eyes Pigmentation 39
Discussion 40
Conclusion 42
Chapter 3 Measuring European Population Stratification with Microarray Genotype Data 43
Introduction 44
Methods and Material 44
Population Samples 44
Detecting Relatedness 46
Network Analysis 47
Principal Coordinate Analysis 47
Bayesian analysis 48
EuroAIMs Selection 48

Measuring Stratification ... 52
Statistical Analysis ... 52
Results and Discussion ... 53
Europe and Neighboring Continents ... 53
Inside Europe ... 56
Stratification in European-derived samples ... 59
Conclusion ... 61
Chapter 4 European Populations and Individuals in Light of 317k SNPs Genotyping ... 64
Introduction ... 65
Methods and Material ... 67
317k genotyping ... 67
Population samples ... 67
Individual samples ... 68
EuroAIM selection ... 70
Results ... 71
Sample Pools ... 71
Individual Genotypes ... 72
EuroAIMs ... 74
Phenotypes ... 77
Functional Genes ... 79
Discussion ... 80
Population pools ... 80
Individual genotyping ... 82
Phenotypes ... 84
Chapter 5 Conclusion ... 86
Methods ... 87
Practical Considerations ... 88
Getting real on human genetic diversity ... 92
Is this racism? ... 93
Final considerations ... 94
REFERENCES ... 96
APPENDIX A Consent Form ... 103
APPENDIX B Coriell Samples Detail ... 108
APPENDIX C DNAPrint EuroDNA 1.0 AIMs ... 109
APPENDIX D Example EuroDNA 1.0 Test Result ... 110
APPENDIX E 180 Individuals Typed on 317k Arrays: Ancestry and Phenotypes ... 111

LIST OF FIGURES

Figure 1.1– A historical snapshot of European migrations. Reproduction from J. Huxley[2] 5

Figure 2.1 – Skin color distribution in Europe, from an adaptation of Biasutti's map of the Old World[60] by Brace and Montagu[61]. 11

Figure 2.2 – Hair color map of Europe by Beals and Hoijer[70] as presented by Frost[58]. 12

Figure 2.3 – Eye color map of Europe by Beals and Hoijer[70] as presented by Frost[58]. 12

Figure 2.4 – Modified Figures from DNAPrint: *A, structure* estimates of parental individuals for EuroDNA 1.0 groups. *B,* Geographic distribution of parental (P) and other population samples from Coriell or from DNAPrint's private collection. 16

Figure 2.5 – PCoA of three artificial populations of eight individuals each. 19

Figure 2.6 – DNAPrint ancestry test results. Each stripe represents an individual. White stripes are missing results. *A,* AncestryByDNA 2.5 (Red=EU, Blue=AF, Yellow=EA, Green=NA) and *B,* EuroDNA 1.0 (Yellow=NE, Red=SE, Green=ME, Blue=SA) 23

Figure 2.7 – Results from *structure* with K = 2 to 7 clusters, using the full set of individuals. The cluster percentages are averaged for each population sample. The width of stripes is proportional to sample size. "Germanic" includes speakers of Germanic languages (mostly Irish, Icelandic, and some Coriell Northern Europeans). "Slavic" includes mostly Russian and Polish individuals. 23

Figure 2.8 – Results from *structure* using the reduced set of individuals. *A,* each stripe representing an individual, and *B,* with cluster percentages averaged for each population, *C,* PCoA. 25

Figure 2.9 – Decay of percentage of the variance explained by each PC for the full set (red) and the reduced set (blue). 27

Figure 2.10 – Top four PCs (reduced set of individuals), with ANOVA correlation coefficients from Table 2.2. 28

Figure 2.11 – Hair pigmentation distribution for *A,* all available individuals, and *B,* European-derived individuals only 29

Figure 2.12 – Hair pigmentation (M-index) distribution for largest European samples 30

Figure 2.13 – Skin pigmentation (M-index) distribution for *A,* all available individuals, and *B,* European-derived individuals only 31

Figure 2.14 – Skin M-index distribution for Europeans and Jewish American samples. 32

Figure 2.15 – Eye luminance for *A,* all individuals and *B,* European-derived individuals 33

Figure 2.16 – Eye luminance distribution for largest European samples and Jewish Americans. 34

Figure 2.17 – Spearman correlation and plotting of hair M-index with genetic ancestry for all individuals with hair data (Table 2.1) 35

Figure 2.18 – Spearman correlation and plotting of hair M-index with genetic ancestry for all European individuals with hair EuroDNA 1.0 results and hair data (Table 2.1). 36

Figure 2.19 – Spearman correlation and plotting of hair M-index with genetic ancestry, keeping only individuals from traditional populations of Europe with hair data (subset from Table 2.1). 37

Figure 2.20 – Spearman correlation and plotting of skin M-index *A,* with worldwide genetic ancestry using world populations and *B,* with European ancestry and individuals. 38

Figure 2.21 – Spearman correlation and plotting of Eye luminance *A,* with worldwide genetic ancestry using world populations and *B,* with European ancestry and individuals. 39

Figure 2.22 – EuroDNA1.0 NE and SE distributions on selected Europeans from the North (Dutch, English, Icelandic, Irish, Russian, Swiss, Polish, Russian, Danish, German, Norwegian, Lithuanian) and South (Armenian, Bulgarian, Greek, Spanish, Italian, Portuguese). 41

Figure 3.1 – Stability of the PC1-PC2 distribution of European individuals across marker sets with different minimum inter-marker separation (50Kb and 100kb). 47

Figure 3.2 – Population structure in European individuals. *A,* Geographically oriented PC 1 and 2 putting forward the 3-cluster model (Model-1) and *B,* adding PC 3 for a 3D representation of 5 putative clusters for Model-2. PCoA is based on average inter-individual ASD over 9,114 SNPs. PC 1, 2, 3 and 4 (Figure 3.3) respectively explain 2.05%, 1.7%, 1.6% and 1.5% of the variation and were highly significant by SKT and ANOVA testing (Table 3.1). *C,* Bayesian clustering analysis using *structure* with the same markers and European individuals..... 50

Figure 3.3 – PCoA plots for the first six PCs in European samples. The bolded vertical bar represents the median PC value of each group, the two hinges are the first and third quartile and notches give an approximate 95% confidence interval for the difference in two medians. Adjusted R^2 is from the ANOVA test. 51

Figure 3.4 – Distribution of F_{ST} between northern (n=36) and southeastern (n=31) cohorts of individuals selected from PC1 values in Figure 3.2 (above 0.5 or below -0.5). *A,* Histogram using all 9,721 SNPs available. *B,* Plot of top 50 SNPs of highest F_{ST} (see also Table 3.4). 51

Figure 3.5 — Population structure in European, African and Asian individuals. *A and B,* PCoA results based on average inter-individual ASD using 9,100 SNPs. PC 1, 2 and 3 respectively explain 11.6%, 3.4% and 1.4% of the variation. *C,* Bayesian clustering results using *structure*[43,44] with the same markers and individuals. 54

Figure 3.6 – PCoA box plots for the first six PCs in samples from Europe and neighboring continents. The bolded vertical bar represents the median PC value of each group, the two hinges are the first and third quartile and notches give an approximate 95% confidence interval for the difference in two medians. The overall correlation between group membership and PC value is

reported by ANOVA's adjusted R^2 for each PC. The few subsequent PCs which are also significant pertain to Europe and are best observed in Figure 3.2 and 3.5. 55

Figure 3.7 – Population structure in panels of European-derived ancestry within the context of European individuals (from Figure 3.2A). *A*, PCoA of the Coriell "Caucasian" panel (n=42) together with Europeans of known ancestry, based on all 9,114 SNPs in common. *B,* PCoA of the CEPH Utah individuals (n=74) and Europeans using all 6,207 SNPs in common. *C, structure* runs using the Coriell "Caucasian" sample based on the full SNP dataset (bottom) as well as sets of different number of north-southeast EuroAIMs (Appendix C). 58

Figure 4.1 – Population samples genotyped on Illumina 317 mapping arrays. Pool samples are indicated in yellow circles with the sample abbreviations from Table 4.1. Sampling sites of genotyped individuals are represented in orange circle; IR=Dublin (Ireland), IT=Rome (Italy), PL=Warsaw (Poland), PT=Porto (Portugal), *A,* Europe and North Africa, *B,* Eastern Europe and Central Asia. 66

Figure 4.2 – PCA genotype data of 26 *pooled* population samples from Europe, Central Asia and North Africa. See Table 4.1 for 3-letter codes, sample sizes and provenance details. Each of the two German, Russian and Basque pools are merged. *A,* Percentage of the variance explained by each PC, *B,* Top two PCs reflecting geographic distribution, *C,* Top 10 PCs, *D,* Zooming in from plot B on the cluster of European pools. 71

Figure 4.3 – PCoA of all 180 *individuals* (Table 4.2), using the full 317k SNP panel. *A,* Decay of the percentage of the variance explained by top 10 PCs. *B,* PC1 with individuals of specific ancestry. Abbreviations correspond to ethnicities listed in Table 4.2 except for Italians sub-groups (ItalianS=South, ItalianC=Center, ItalianCN=Center-North) and island individuals (Sard=Sardinian, Sic=Sicilian). *C,* Top 3 PCs combined. Letters refer to individuals listed in Table 4.2. Irish are in green, Polish in red, Portuguese in black, ItalianCN in orange, ItalianC in blue, ItalianS in cyan. *D,* PC1 with non-European axis PC4. 73

Figure 4.4 – Average heterozygosity for the four samples (leaving out foreign and admixed individuals, Table 4.2). 74

Figure 4.5 – PCoA of all 180 *individuals* (Table 4.2), using AIM panels. Abbreviations correspond to ethnicities listed in Table 4.2 except for Italians sub-groups (ItalianS=South, ItalianC=Center, ItalianCN=Center-North) and island individuals (Sard=Sardinian, Sic=Sicilian). Irish are in green, Polish in red, Portuguese in black, ItalianCN in orange, ItalianC in blue, ItalianS in cyan. *A and B,* PCoA with 145,745 AIMs of F_{ST} >0.01 among 21 European pools (Figure 4.2D). *C and D*, Using SNPs with F_{ST}>0.03 among the pools and keeping the 12,035 SNPs of null F_{ST} among the 4 groups (AIM panel used in Figure 4.6). 75

Figure 4.6 – Bayesian analysis with *structure* using ~12k AIMs (Figure 4.4C). Country labels represent autochthonous individuals from the four corresponding sampling sites. Foreign individuals are represented by letters a-v in the same order as in Table 4.2. 76

Figure 4.7 – Phenotype inter-correlations. 78

LIST OF TABLES

Table 2.1 – Available subjects: results and data for each population sample..... 14

Table 2.2 – PCoA significance using SKT and ANOVA test. 26

Table 2.3 – Differences between hair pigmentation distributions 29

Table 2.4 – Differences between skin pigmentation distributions..... 31

Table 2.5 – Differences between eye luminance distributions 33

Table 3.1 – Individual samples description 45

Table 3.2 – PCoA significance tests 46

Table 3.3 – European stratification models: Correlation between group membership and individual PC. 48

Table 3.4 – Top 20 northern-southeastern EuroAIMs (Model-0)..... 57

Table 3.5 - Model-1, SKT Combined P-values for PC1 (next 5PCs never significant). F_{ST} cutoff of 0.10 yields between 269 and 380 AIMs per cluster pair. 59

Table 3.6 - Model-2, SKT Combined P-values for PC1 (next 5PCs never significant). F_{ST} cutoff of 0.15 yields between 158 and 833 AIMs per cluster pair. 60

Table 4.1 – Population samples pooled on the Illumina 317k mapping array..... 68

Table 4.2 – 180 European individuals, genotyped on the Illumina 317k mapping array 69

Table 4.3 – Phenotype inter-correlations of Irish, Italian, Polish and Portuguese samples..... 77

Table 4.4 – Pigmentation candidate genes and SNPs with highest pairwise F_{ST} among the 4 European groups (individuals typed on the 317k mapping arrays) 79

Table 4.5 –Mean $F_{ST} \times 10^3$, using all 317k SNPs 82

Table 5.1 – Genetic ancestry tests for Europeans 89

ACKNOWLEDGEMENTS

I am grateful to my advisor Mark Shriver whose intellectual stimulation and generosity carried me through this work. I thank the members of my committee for thoughtful suggestions and comments. The many people of the Penn State Anthropology department, past and present, provided an inspiring environment. This work would not have been possible without all students and collaborators in the US and Europe who helped and participated in sample collection. In particular Brian McEvoy, Sandra Beleza and Jun Li provided assistance and stimulating intellectual collaboration; Laurel Pearson and Ellen Quillen provided invaluable proofreading and genotyping. This research was funded in part by Hill and Baker grants, and by grants to Mark D. Shriver from the NIH/NHGRI (#HG002154) and from the Science Foundation of Ireland (Walton fellowship 04/W4/B643). The Weiss Fellowship funded my living expenses and tuition.

Dearest friends and family, you know who you are; you know I am deeply grateful.

Chapter 1

Introduction

European Population

Among continents Europe is remarkable for its relative small size, dearth of migration barriers and abundance of historical population movements[1,2]. These features suggest low levels of genetic differentiation and stratification in Europe. Indeed, compared to other continents, Europe has by far the smallest inter-population genetic distances[3], a relative lack of population structure and a genetically homogeneous pattern extending to North Africa, West Asia and beyond[4,5]. Johann Friedrich Blumenbach (1752-1840) proposed the concept of a "Caucasian race" as the central one of five "varieties of mankind", deriving the name from the Caucasus region because in his words it "*produces the most beautiful race of men*" and "*in that region, if anywhere, it seems we ought with the greatest probability to place the autochthones of mankind*"[6]. After two centuries of drastic revisions, anthropology has distanced itself from this scheme. However the labels "Caucasian" or "White", used to characterize individuals who "look European" still persist. Typical physical traits usually used to characterize these individuals are neither unique to Europe nor fixed within any large European population (e.g. blond hair, light eyes and skin). As a result, the concept of a "Caucasian" biological entity has lead to much scientific, social and legal confusion[7], but it is still commonly used by laypersons and some scientists—in particular in the social sciences, psychology and medical research, where the term "Caucasian" may take on different meanings. In the end, is it biologically justified to assume the existence of a biological "Caucasian" entity, different enough from other world population to deserve such a subspecies-like denomination? Do the data support the conclusion that genetic variation within Europe is of no use or significance to biomedical research[8]? More generally, can we characterize or measure any such genetic difference at any population level? In light of historical considerations, is it ethical to ask and pursue such questions?

Before reviewing the latest genetic evidence and presenting new results, I must define other variables of importance in the analysis of the genetic data presented here. Part of the problem surrounding concepts such as a "Caucasian race" boils down to defining what the term "population" means, in particular when applied to humans. Mathematically a population is simply a set of things, here humans. But how do we define which humans are of interest to group together? And why would we want to do that anyway? Even though all human groups throughout the world are inter-fertile, in practice the human species is not an ideal panmictic population. The first obvious reason is geographical distance, although this situation alone may only generate smooth clines of genetic variation. Geographical barriers are one element generating relative genetic isolation and differentiation. However, populations can sometimes also seem differentiated even after long term cohabitation in the same region. Indeed, culture differences (mainly along religious and linguistic

lines) provide another important and complex barrier to gene flow among human groups. Groups of individuals separated from one another physically or culturally are reasonably prone to reproductive isolation. Over time, the processes of genetic drift, unique mutations, differential sexual selection and natural selection in a variety of environments are unavoidable; they result in measurable genetic differentiation among human groups. Indeed, using classical gene frequency differences, linguistic boundaries have been shown to correlate with zones of sharp genetic change[9,10]. Different statistical analyses on similar data[11] has also shown abrupt gradients of variation, although the fault lines did not necessarily correlate with language. In the present thesis, ethno-linguistic labels are assigned to individuals and populations based on all available information on their origins; however analyses are performed on the genetic data without *a priori* assumption on individuals' cultural affiliations. In return it is possible to test the relevance of those ethno-linguistic labels. For instance as we will see below, individuals from culturally homogenous nations which may be considered populations (e.g. France or Italy) are not guaranteed to show genetic homogeneity, and sub-population labels may need to be applied accordingly. This is one critical advantage of using individuals as the statistical unit rather than population samples.

Despite their extensive geographical range, human populations have not become separate phylogenetic entities, i.e. subspecies, thanks to recent common ancestry and possibly long term inter-population gene flow—especially in Europe. Therefore sharp genetic distinctions such as fixed allelic differences between populations are likely be rare or absent, and inferring past migrations and population events from genetics alone may be hazardous at best. Archaeological and historical records point to many instances of both isolation and gene flow in the past ~40,000 years of human occupation of Europe[1,12], but the interpretations are potentially endless because we do not usually know either the pre-migration genetic structure of populations involved, or the extent to which cultural transmission proceeded without extensive migration. Despite these limitations it is instructive to review the major known demographic events of Europe as a population and, possibly, a set of populations.

Europe before anatomically modern *Homo sapiens*

The archeological record shows that the genus *Homo* first appeared in Europe at least 800,000 years ago and it is still a matter of debate how many *Homo* species and subspecies roamed the continent until the rise of the classic Neandertals from European *Homo heidelbergensis* between 300 and 200,000 years ago[13]. Our species, anatomically modern *Homo sapiens* (amHs), also evolved from *Homo heidelbergensis*-like ancestors, most likely in Africa more than 100,000 years ago, before entering Europe by 40,000 years ago[14,15]. A consensus is growing that Neandertals contributed nothing

or very little to amHs gene pool[16-18], although both species cohabited for about 10,000 years before Neandertals vanished 30,000 years ago[19]. However some genetic evidence still fuel a debate that Neandertals may have contributed a significant amount of genes to amHs, even possibly selectively advantageous variants[20,21].

The Neolithic revolution

Genetic differentiation among contemporary European populations has been represented as gradients, mainly following a southeast to northwest cline; in the light of archeological data this trend was interpreted as population expansion following the rise of agriculture and sedentism during the Neolithic[11,22-24]. It is generally accepted that the history of European populations was strongly influenced by the 'Neolithic Revolution', when a range of technological improvements including animal and plant domestication spread from the Near East some 10,000 years ago[1,25]. This is reflected in that most European languages spoken today are of the Indo-European language family which also originated in the Near East or Anatolia and is thought to have expanded from there, along with plant and animal domestication[26]. However the extent to which genes traveled with the Neolithic lifestyle remains unresolved. On one end of this debate is the Demic Diffusion (DD) model, which advocates that agriculturalists replaced local hunter-gatherer populations with little or no genetic intermixing[23]. On the other end, the Cultural Diffusion (CD) model has been opposed, claiming that the Neolithic expanded mostly through cultural transmission. The CD model is supported on the grounds of computer simulations showing that the data are consistent with a large genetic contribution of Paleolithic and Mesolithic populations to the European gene pool, and that the same clines are equally likely under both CD and DD models[27]. The potential role of Mesolithic re-expansion from glacial refugia is recognized but downplayed in the current literature[28] and has not been formally tested in simulations. Similarly the pre-Neolithic genetic structure in Europe has not been considered, although it may be large considering the area was inhabited for over 30,000 years by scattered groups hunter-gatherers and horticulturalists. In reality, nobody today seriously advocates a 100% CD or DD model, as the debate has shifted to admixture levels[14,29]. Various estimates based on mitochondrial DNA (mtDNA) and non-recombining Y-chromosome (NRY) data have yielded estimates from 10% to 50% Neolithic admixture in the European gene pool, the rest purported to be of Paleolithic origin[30,31]. But the confidence intervals are large and admixture estimates differ widely according to regions. Indeed, despite highly sophisticated methods, two markers (mtDNA and Y) or several dozen classical markers are limited in statistical power relatively to genome-wide markers, such as the thousands of Single Nucleotide Polymorphisms (SNPs) which analysis I present in the following chapters. Finally, even

under the unlikely scenario where original parental populations are as clear-cut as Paleolithic vs. Neolithic (hunter-gatherers vs. farmers), it remains to be seen how much the intricate post-Neolithic migrations (e.g. Figure 1.1) may have muddled the story beyond recognition.

More accurate inferences on the peopling of Europe may become possible by increasing the number of genetic markers by several orders of magnitude with microarray technologies.

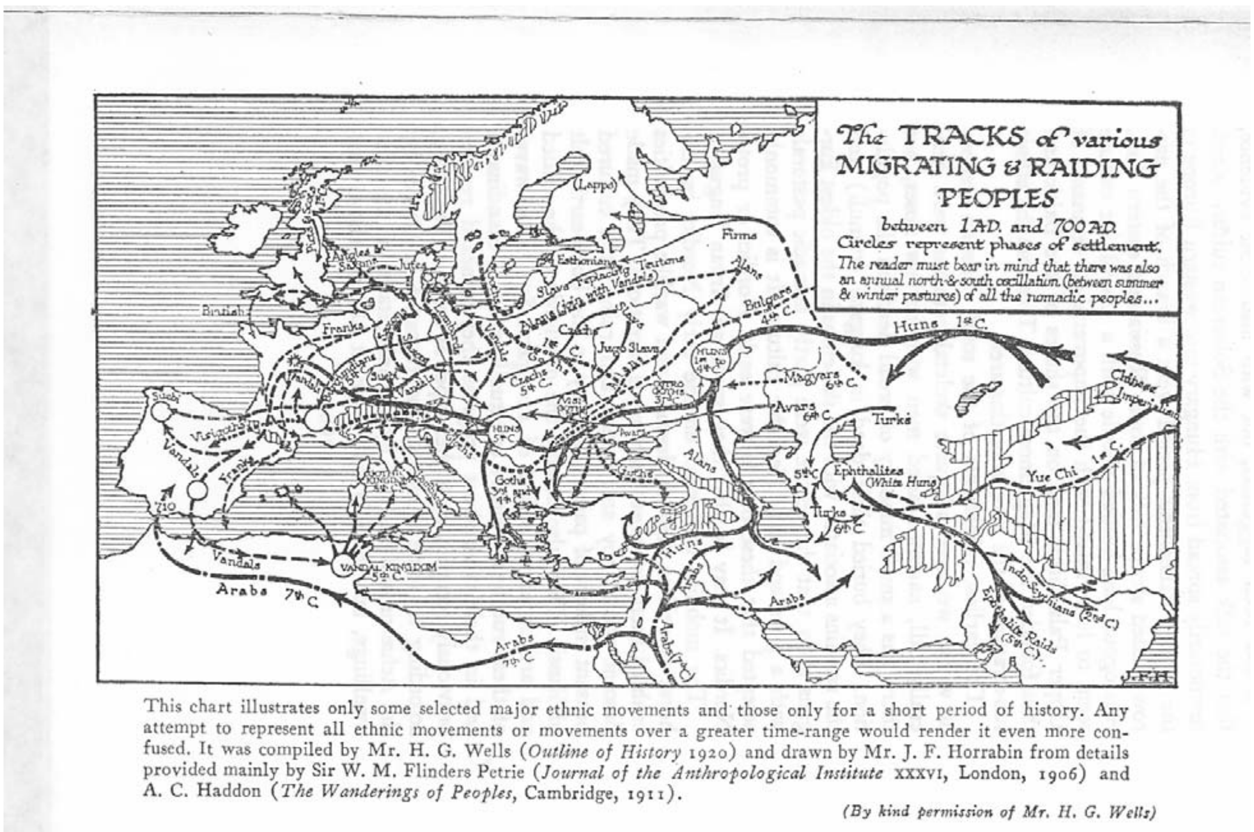


Figure 1.1– A historical snapshot of European migrations. Reproduction from J. Huxley[2]

Genetic Markers

In an analysis using 1,915 world population samples with 120 protein polymorphisms, Europeans displayed the smallest inter-population distances and close relationships to North African and West Asian populations[3]. Using first 377, then 933 microsatellite markers typed on the 52 worldwide HGDP-CEPH populations (~1,000 individuals total), the European and neighboring populations from the Middle East, South and Central Asia have been shown to form a genetic cluster at the worldwide level[4,32]. Recent analyses using between ~6,000 and 10,000 genomewide SNPs have shown significant patterns of structure in Europe[33] and Eurasia[34] albeit on a limited number of populations.

Chapter 2 will focus on the 313 Ancestry Informative Markers (AIMs) of a commercial genetic test, EuroDNA1.0 from DNAPrint (Sarasota, FL), typed in large numbers of individuals from Europe and neighboring populations. Chapter 3 will present the analysis of genome-wide typing of *ca.* 10,000 single nucleotide polymorphisms (SNPs) from Affymetrix (Santa Clara CA) 10k mapping arrays in 297 individuals[35]. Chapter 4 will present the analysis of data from pools of European samples as well as individuals genotyped for *ca.* 317,000 SNPs on the 317k Illumina (San Diego, CA) mapping arrays.

Methods

Traditional linguistic and geopolitical fault lines in Europe do not exactly match and it is not clear which, if either, correlate best with potential genetic structure. Despite decades of efforts using available physical, physiological and partial genetic data, the bulk of the genetic landscape of human variation has just begun to emerge in the past few years. Recent studies involving large numbers of SNPs demonstrated population stratification within European American samples which could not be detected with standard methods[36-38], hence the need to seek and apply new ways of detecting stratification and screening for AIMs to control for it genetic association studies[39,40].

I first applied methods excluding any *a priori* ethnic information on the individuals genotyped, namely principal coordinate analysis (PCoA[34,41,42]) and Bayesian Analysis with the program *structure*[43,44]. I used two main methods of measuring the significance of axes of stratification. The first method is a variant of split-half reliability testing[45]. It consists in splitting the marker set in two independent halves, using each half to calculate ancestry components, and measuring the correlation between individual results obtained with each marker set. Significant correlation shows that the measure of ancestry is reliable for the axis of concern. The second method measures how group membership correlates with the various statistics measuring genetic variation in these analyses. Although often yielding similar results, the two methods actually measure slightly different concepts. The first is a measure of reliability, whereas the second measures the significance of observed separations between ethno-linguistic groupings. A measure of the actual level of stratification between two ancestry components is presented in chapter 3, through the worst-case scenario of a gene association study where all cases and controls entirely come from different populations. I simulated 1,000 individuals for each cohort based on the measured allele frequencies; the mean chi-square distribution allows estimating the maximum confounding which the observed genetic stratification would generate in a case control study[39]. Finally, standard statistical methods were used to test ancestry-phenotype and phenotype-phenotype associations (chapters 2 and 4).

Phenotypes

The focus of chapter 2 and 4 will be another important set of variables, the human pigmentation traits measured for most of the individuals genotyped. Melanin content was used to measure hair and skin pigmentation. The qualitative and quantitative assessments of eye color will be described in chapter 2. Although they are more easily defined, pigmentation traits are directly related to population concepts and may or may not correlate with cultural population labels and AIMs. But discovering the genetic basis for those traits has obvious application in biomedical research, forensics and anthropology. The evolutionary mechanisms underlying pigmentation traits also have an important educational value; instead of continuing to be racially divisive these traits must be celebrated as vestiges of our evolutionary journeys (chapter 5).

Stratification

Genome-wide association studies are becoming a key tool in attempts to map genes underlying many complex traits. However, the presence of population stratification, or individual ancestry differences within and among samples, may jeopardize such approaches. In particular, discordant ancestry levels between cases and controls can lead to false positive association with a trait and/or reduced power to detect such associations. The issue is most acute and widely recognized in individuals who differ in continental origin, or who are admixed between such populations. Ancestry Informative Markers (AIMs), typically SNPs, that show large frequency differences between inter-continental groups, can be used to detect and correct for such stratification. However, the study of intra-continental structure is less well explored or understood.

Because of the low apparent diversity in Europe, it has been argued that European population stratification does not represent a significant source of bias in epidemiological studies[8]. However recent autosomal SNP studies have highlighted significant patterns of structure within Europe along a North-South axis[46]. The potential confounding influence of this stratification on association mapping studies in European-derived population samples was also recently demonstrated[36,37]. Beyond these first insights, little is known about the geographical distribution and complexity of European genomic structure. The identification of additional significant patterns likely requires more extensive population samples and greater numbers of markers such as will be presented in chapters 2 and 4. One purpose of the present work is to look for the most genetically meaningful genetic structure in Europe, sampling with an attempt to represent the broadest ethno-linguistic diversity as first hints to potential genetic stratification. I will present several population models in chapter 2 and 4, which may serve as starting point for further refinements.

Admixture v. clines

The term 'genetic admixture' generally refers to the intermixing between human parental populations, defined as relatively homogenous groups with high enough genetic distance among each other that admixture proportions may be measured. This can only be possible if markers exist that show significant allele frequency differences among populations. Such ancestry informative markers (AIMs) have been screened to measure admixture among any or all populations of Europeans, West Africans, Indigenous Americans and East Asians[47,48], resulting in the first genomewide genetic admixture test[49] by the company DNAPrint (Sarasota FL).

The pattern of worldwide human genetic diversity has in the past alternatively been described as discrete entities[50] or as a huge melting pot with insignificant regional differences[51]. The latter view has gain almost common wisdom with the often unquoted assertion that for the most part genetic variability is contained within rather that among populations. Similarly, more recent studies have emphasized to various degrees discontinuities[9] or the gradual continuity of regional differences[11,52,53]. Similarly when individuals are the unit of representation, a lively debate still opposes whether the best representation is by clines or gradients[54] or clusters[5], and how much it matters for which problems or traits under investigation[4,55]. Indeed the question becomes a mere terminology issue when facing practical aspects of admixture (chapter 2) and genetic stratification as measured in chapters 3 and 4; in fact, as we will see the number and type of markers used makes a huge difference in how we see the apportionment of human genetic diversity. An equally important aspect of this debate is how this scientific knowledge of human genetic diversity is perceived in the public, which I will address in chapter 5 (conclusion).

Chapter 2

European Genetic Ancestry and Pigmentation Phenotypes

Introduction

From insects to birds and mammals, many animals use the pigmentation of their conspecifics as clues to social interactions, including mating decisions, and humans are no exceptions. In many times and places, pigmentation variations in skin have been and are still used as proxy to social oppression[7], although originally the geographical variation in skin pigmentation is strongly determined by natural selection forces under different climates[56]. Additionally, skin, eye and hair color are major factors in human mating and are therefore likely to be shaped by sexual selection[57,58]. Human pigmentation traits have been used by scientists as part of the battery of visible and measurable phenotypes useful to describe and investigate human variation, from antiquity to the 18th century[6]. The focus of the present chapter will be the relation between European distribution of skin, hair and eyes pigmentation and genetic ancestry measured by genetic markers of a commercial test, EuroDNA 1.0™ (DNAPrint, Sarasota, FL).

Europe is the region of lowest skin pigmentation in the world. Despite problematic measurement methods[59] Biasutti's data and maps of skin pigmentation dating back to the 1950's[60] are still the reference in most textbooks and scientific publications on the subject. I present here the European part of an improved version[61] showing that the commonly accepted north-south pattern is complemented with an east-west component (Figure 2.1). This pattern of concentric clines around the south Baltic region is accentuated in head hair and eye color maps (Figures 2.2 and 2.3). Indeed pigmentation of hair and eyes not only is some of the lightest in the world, but also displays the highest variation in Europe[58]. However this variation is not random and all three traits are inter-correlated as shown by the three color maps (Figures 2.1, 2.2 and 2.3) as well as in a survey of the Danish population[62] and in Chapter 4.

Although tremendous progress has been accomplished on the genetic basis of human pigmentation traits[63,64], and recent studies have demonstrated links between genetic ancestry and the pigmentation of skin[47,65,66], hair[67] and eyes[68], the inter-population variation patterns remain largely to be described and compared with genetic ancestry.

Genetic ancestry tests are of two main types: those which rely on uniparental DNA (NRY and mtDNA) and those which use a panel of genome-wide autosomal markers, typically SNPs[49,69]. These SNPs are chosen for being Ancestry Informative Markers (AIMs), i.e. they are characterized by high allele frequency differences among samples from putative parental populations. In this chapter I evaluate the usefulness and relevance of the AIM-based ancestry test EuroDNA 1.0, which is the only genomewide test of European ancestry currently on the market. After reviewing the functional aspects

of the markers used in this test, I will present the evaluation of EuroDNA1.0 ancestry profiles in 530 individuals of Europe and peripheral world populations. The individual genotypes of the 313 SNP markers of EuroDNA1.0 were analyzed with various statistical methods. I will then present the patterns of correlation between pigmentation traits (skin, hair and eyes) and three measures of ancestry for Europe (EuroDNA1.0, self-defined ethnicity and genetic distances).

Although successful as a commercial test, EuroDNA 1.0 remains to be evaluated in a context of large samples of individuals from Europe and its periphery. I will present here such an evaluation both from the perspective of EuroDNA 1.0 ancestry test results as well as analyzing directly the raw genotype data of EuroDNA 1.0 AIMs. Finally, the possible correlation between various physical traits and ancestry proportions from genetic tests will be presented.

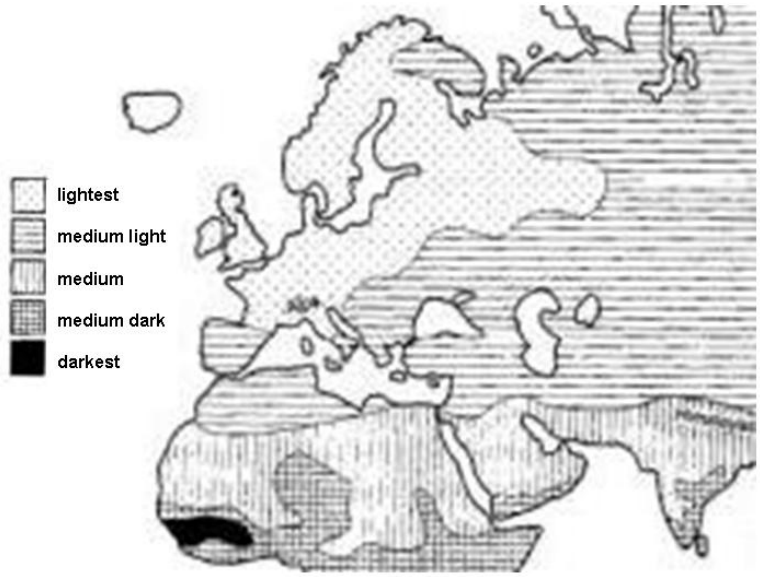


Figure 2.1 – Skin color distribution in Europe, from an adaptation of Biasutti's map of the Old World[60] by Brace and Montagu[61].

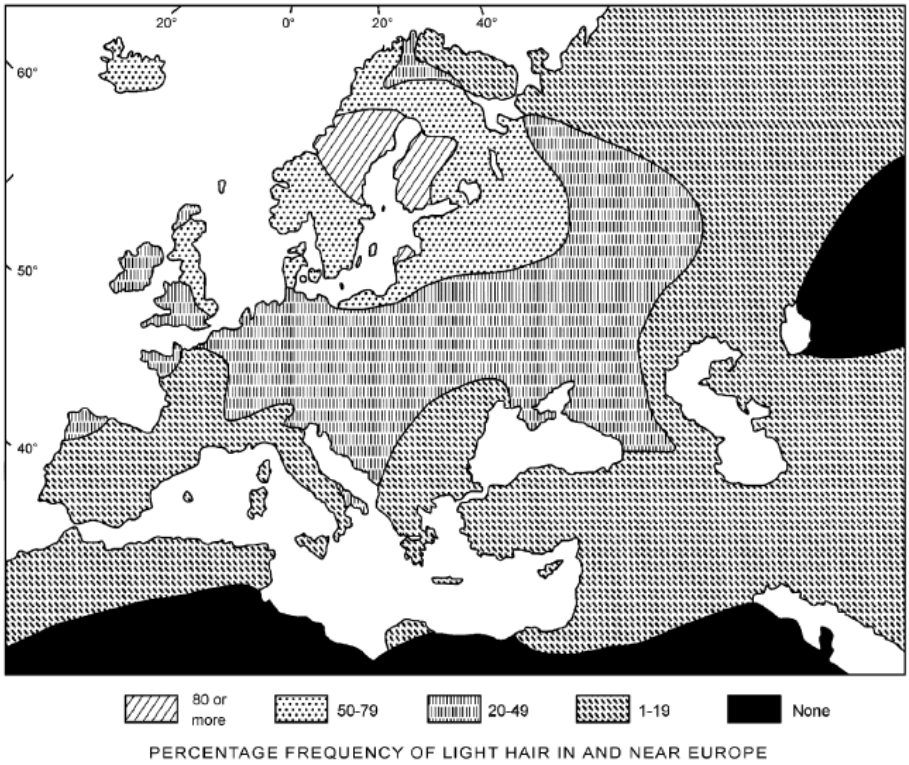


Figure 2.2 – Hair color map of Europe by Beals and Hoijer[70] as presented by Frost[58].

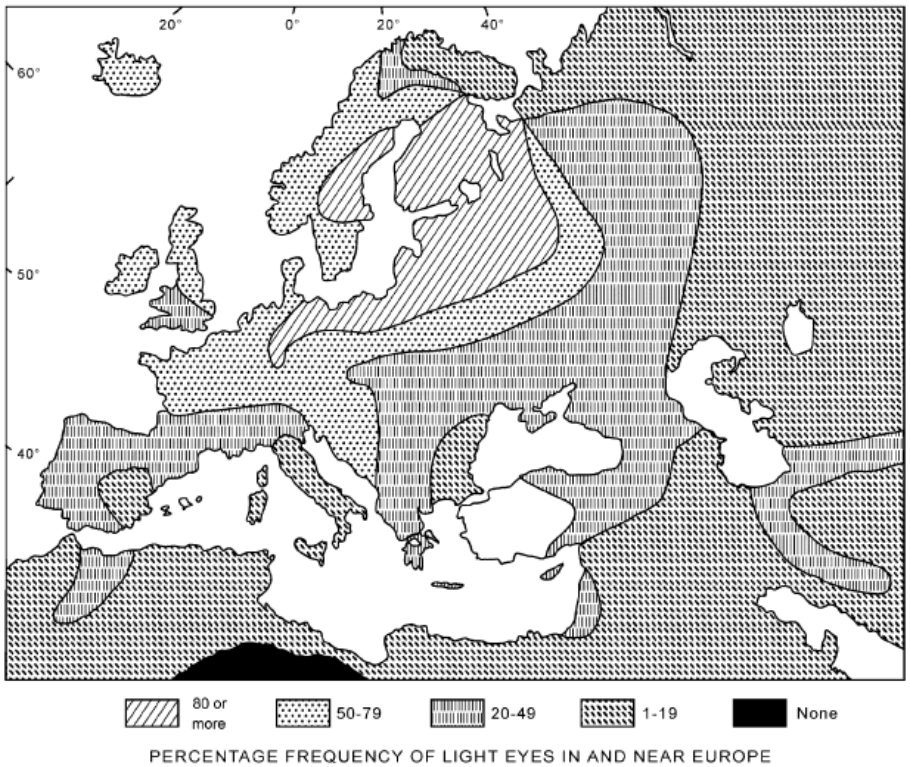


Figure 2.3 – Eye color map of Europe by Beals and Hoijer[70] as presented by Frost[58].

Material and Methods

Population Sampling

From the summer of 2005 to the summer of 2006, I coordinated data collection at Penn State under Dr. Shriver's supervision and with the help of the other students from his lab. Dr. Shriver and I wrote and submitted the IRB protocol and consent form for the "Genetics of Human Pigmentation and Facial Features" project, as well as regular updates (latest version at the time: Appendix A). The first sampling session in June 2005 involved mostly European citizens which I recruited in the European club at Penn State (n=37 persons). Then in the fall 2005 and spring 2006, 1,394 subjects were sampled, mainly US citizens from Dr. Richard's "Race and ethnic relations" class at Penn State (SOC119) in collaboration with him and his team. A few participants from the Middle East were also recruited from Penn State students clubs. Human DNA samples from Europe were collected in several steps. I first went to France in July 2005, where I collected DNA samples in three different regions under a modified protocol to that used for the Penn State collections, primarily having limited phenotype assessment (n=93). Then in 2006, Dr. Mark Shriver, Dr. Brian McEvoy (Trinity College, Dublin) and I visited five major European cities and with the help of local collaborators collected DNA samples and all the complete battery of phenotypes as in the Penn State sessions. The five cities were Paris (France), Rome (Italy), Warsaw (Poland), Porto (Portugal) and Dublin (Ireland), where DNA and phenotypes were collected from n=76, 137, 102, 202 and 203 individuals respectively. Participants were mainly from the region where they were sampled and the declared ancestry of their four grandparents along with other demographic information was recorded. Before each sampling session a presentation was given by me, Dr. Shriver or B. McEvoy to the potential participants, mainly describing the content and meaning of the consent form (Appendix A). All participants (save the Polish and Irish volunteers) were given AncestryByDNA 2.5 ^{TM} (DNAPrint, Sarasota, FL) test results as for donating DNA and being measured for a variety of phenotypes, including pigmentation measure of hair and skin and eye photos (Appendix A). The earlier European and French participants in 2005 were also given their EuroDNA 1.0 results as incentive.

The commercial population samples from the Coriell Institute didn't contain any phenotypic information. Personal information on sample origins is limited, with sample names, ethnicity and additional remarks sometimes in contradiction (for a detailed description see Appendix B).

Table 2.1 – Available subjects: results and data for each population sample

Sample Label *	All	ABD2.5	EUR1.0	Hair	Skin	Eyes
AfrAm	84	66		44	74	39
Arab	15	12	12	6	6	3
Arab (ME)	8	8	8			
Armenian	1	1		1	1	
AsianMix	6	4		3	4	3
Basque	10	10	10			
Bulgarian	4	4	2	4	4	2
Caribbean	61	27		32	46	19
CentralAsian	5	4	3	2	3	2
Chinese	30	23		11	14	8
Dutch	1	1			1	1
Dutch (NE)	1	1	1			
EastAfr	7	7	5	5	7	4
English	3	1	1	2	2	2
EuroAfr	7	5		6	6	4
EuroAm	925	841	16	498	846	488
EuroAsian	11	6		6	7	4
EuroMix	36	28	11	27	36	34
EuroMix (NE)	1	1	1			
EuroNat	74	65	1	33	65	48
French	169	167	160	25	75	70
French (NE)	2	2	2			
Georgian	1	1	1	1	1	
German	5			5	5	5
German (NE)	1	1	1			
Greek	2	1		2	2	2
Greek (SE)	10	8	8			
Hungarian	2	1		2	2	1
Icelandic	13	10	13			
Iranian	5	5	5	5	5	3
Iranian (ME)	2	2	2			
Irish	147	1		123	146	145
Irish (NE)	1	1	1			
Italian	140	139	77	85	130	123
Japanese	10	10				
Jewish	39	37	33	32	37	21
Korean	17	16		10	16	7
LatinAm	52	40	1	27	43	16
Melanesian	5	5				
NewMix	145	129	4	91	135	81
NorthEuroMix	37	11	6	28	36	36
NorthEuroMix (NE)	2	2	2			
Norwegian (NE)	2	2	2			
Polish	84	51	31	63	84	83
Polynesian	2	2		1	1	1
Portuguese	187	187	19	154	184	184
Russian	4	3	3	3	4	3
Scottish	2			1	2	2
SouthEastAsian	23	23		1	3	1
Serbian	3	3	1		2	1
SouthAsian	25	20	12	18	23	9
SouthAsian (SA)	34	34	34			
SouthEuroMix	4	4		4	4	4
Spanish	7	6	4	3	3	3
Swiss	2	1	1	1	1	1
Turkish	3	2	2	3	3	3
Turkish (SE)	34	34	34			
Unknown	14			5	7	
WestAfrican	13	11		5	11	8

Note. — * EuroDNA 1.0 parental individuals; NE=Northern European, SE=Southern European, ME=Middle Eastern, SA=South Asian. Afr=African, Am=American, EuroNat=EuroAm claiming any level of Indigenous American ancestry, AsiaMix=Mixed East Asians, NewMix=Individuals mixed from various world populations, etc.

Eliminating Outliers and Relatives

Individuals of mixed ancestry from two or more European regions were excluded in order to focus on the main genetic differences among these regions. Two Spanish Iberians from Coriell (DNA IDs: NA04340 and NA03780) and two Portuguese of our collection were removed due to high levels of non-European ancestry (detected by the AncestryByDNA 2.5 test, described below). Relatedness among individuals was usually indicated in the questionnaires given to the participants. However I performed ML-Relate analysis as in the previous chapter[71], using all 313 AIMs, in case some relatedness had not been indicated. Finally the twelve individuals with >8% genotyping error were also excluded.

The EuroDNA 1.0 Test

Unlike mtDNA or NRY based tests which use uniparental markers for database matching, EuroDNA 1.0 uses maximum likelihood estimation (MLE) to calculate ancestry proportions based on genome-wide autosomal SNPs selected for ancestry informativeness. In the final version of EuroDNA 1.0 these AIMs add up to 313 SNPs, which includes all 176 AIMs of the AncestryByDNA 2.5 test. The full list and details are presented in Appendix C, where gene information was collated with the *snpper* online database search (http://snpper.chip.org/), based on dbSNP (hg17 build 123).

The AncestryByDNA 2.5 test, DNAPrint's worldwide genetic ancestry test is based on 176 AIMs selected for their ability to differentiate (allele frequency differences) four continental populations: European (EU), Sub-Saharan African (AF), Native American (NA) and East Asia (EA)—which could be more exactly or appropriately named, respectively; Western European, West African, Indigenous American. This was used with success in forensic applications[72] and by a large number of people curious about their ancestry. In most cases markers of AncestryByDNA 2.5 provide a good way to predict a person's genetic ancestry in the framework of the four-population model (EU, AF, NA and EA)[73].

The EuroDNA 1.0 test was later created from a model of 4 parental groups for Europeans (Figure 2.4) named by DNAPrint; Northern European, Southeastern Europe (Mediterranean), Middle Eastern, South Asian. These components are respectively abbreviated as NOR, MED, MIDEA and SA on DNAPrint's online manual (http://www.ancestrybydna.com/welcome/productsandservices/eurodna/ancestrykit). Alternative shortcuts, respectively NE, SE, ME and SA, are used in the EuroDNA 1.0 test results given back to customers (see Appendix D for an example). This latter notation will be used in this chapter to refer to

the four EuroDNA 1.0 components, mainly because NE and SE are more accurate descriptions of the components they refer to.

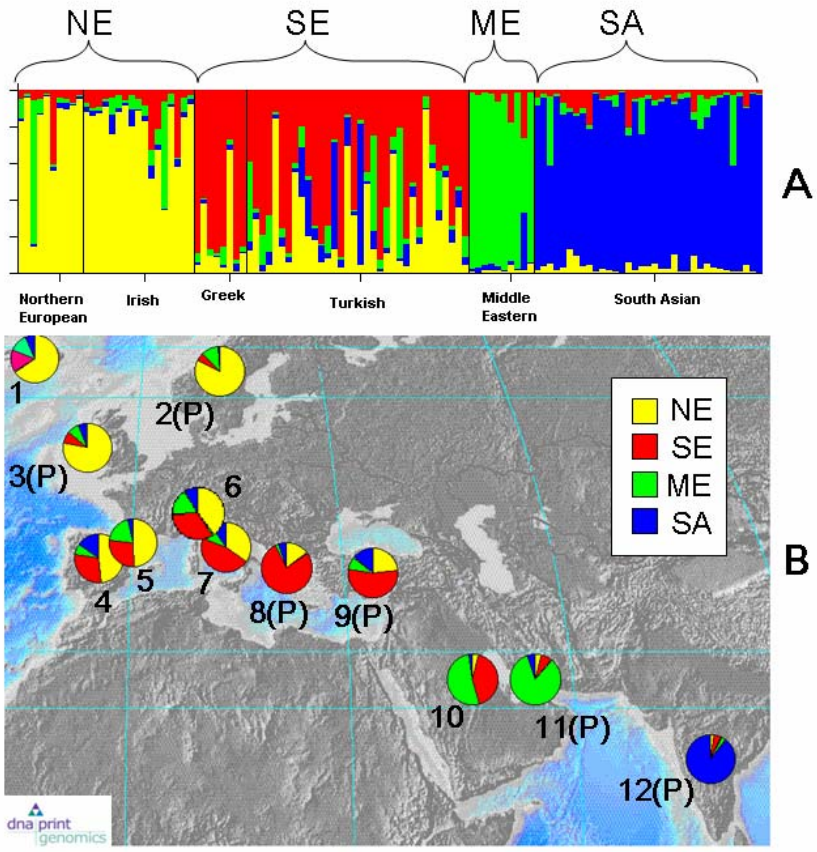


Figure 2.4 – Modified figures from DNAPrint: ***A, structure*** estimates of parental individuals for EuroDNA 1.0 groups. ***B,*** Geographic distribution of parental (**P**) and other population samples from Coriell (*) or from DNAPrint's private collection (**):
1-Icelandic*
2-Northern Europe* (**P**)
3-Irish** (**P**)
4-Iberian*
5-Basque*
6-Ashkenazi Jewish*
7-Italians*
8-Greeks* (**P**)
9-Turkish** (**P**)
10-Middle East 1*
11-Middle East 2* (**P**)
12-South Asian** (**P**)

In addition to the 176 AIMs of AncestryByDNA 2.5, another 125 AIMs were selected by DNAPrint by pooling European DNA of putative parental samples from northern Europe, southern Europe, the Middle East and South Asia on Affymetrix 10K (Santa Clara, CA) mapping arrays. These AIMs were screened for their high allele frequency differences among the 4 groups. All 313 AIMs were then typed in the same individuals and run through *structure*, where K=4 appeared as the best mode. Each of the four ancestry components was defined by "parental" individuals, selected as those who had at least 85% of the corresponding cluster in *structure* (T. Frudakis and I. Halder, personal communication). Most Irish and Coriell Northern Europeans define the NE component; most Greek and Turkish individuals define the SE component; the Coriell Middle Eastern 2 (ME2) sample defines the ME component; most South Asians define the SA component. The Coriell Middle Eastern 1 (ME1) sample is a mixture of Coriell Greeks (used to define the SE component) and Middle Eastern 2 individuals (Appendix B) which explains the mixed appearance on Figure 2.4B. The allele frequencies in the four groups of parental individuals determine the MLE parameters used to calculate an

individual's percentage of affinity to each of the four ancestry components. This individual ancestry is calculated with a DNAPrint proprietary MLE algorithm, which gives the percentages of admixture from each parental population. The probability space is represented by 4 adjacent triangles, i.e. a flattened tetrahedron, with the point of highest likelihood surrounded by probability contours (example in Appendix D, page 2) or a 4-column histogram with likelihood intervals (Appendix D, page 3).

Genetic Distance

The allele sharing distance (ASD) method[74,75] was used to estimate genetic distances between individuals. The genetic distance between two individuals for a given SNP is 0.0 if they have identical genotypes, 0.5 if they share one allele and 1.0 if they have no allele in common. Overall ASD between two individuals was calculated by averaging these distances over all SNPs where they both had genotypes. The result is a triangular matrix containing the ASD of each person compared to all other persons

To measure genetic distance among groups of individuals I used F_{ST} which is based on allele frequencies differences[76]. I used Weir's unbiased F_{ST} corrected for sample size differences[77].

Principal Coordinate Analysis

Principal Coordinate Analysis[41] (PCoA) summarizes multivariate data sets into trends of maximum variance known as Principal Components (PCs). PCoA was chosen over Principal Component Analysis (PCA) as it was shown to have better power to identify clusters[78] and is more robust to missing genotype data. I used R software's ade4 package[79] to conduct PCoA on the ASD matrices[74,75]. Stability of PC values was evaluated by calculating the average and SD over all the possible runs leaving out one individual at a time. The differences between runs were so minute that I subsequently performed a single PCoA run on all individuals of interest.

PC axis significance

While the amount of variation explained by each PC (i.e. their Eigen value) is an indication of its importance, these proportions are not measures of statistical significance and are typically small for very large numbers of markers as in the present dataset. The shape of the Eigen value decay was plotted as a diagnostic indicating the importance and informativeness of each PC (Figure 2.9). As for PC significance, two independent methods were devised (Table 2.2).

First, I tested for linear correlation of PC axes with population or group membership through an analysis of variance (ANOVA) with each PC as dependent variable and population/group membership

as predictor variable. The correlation between group membership and PC, when present, is often not linear and sometimes the variance appeared to vary significantly among groups (Bartlett test). Therefore the P-value was also computed with the non-parametric Kruskall-Wallis test. Since the level of significance was always the same, therefore only the ANOVA P-value is reported. It is of note that lack of significance with the ANOVA method can reflect either lack of population stratification along the PC being tested, or inadequate population labeling; hence proper labeling of samples is a prerequisite to detect stratification with this method.

The second method, which I refer to as the Split Karyotype Test (SKT), offers the advantage of not relying on population labels[34,42]. The SKT is a form of split-half reliability test used in other fields of science[45,80,81]; here the SNP data is divided into two independent marker panels (i.e. non-syntenic) and PCoA is carried out on each panel separately. Under a null hypothesis of no population structure, no correlation is expected between the PCoA results obtained from each panel of SNPs. The correlation is tested in the SKT by calculating the Spearman correlation coefficient between each individual PC for each SNP panel. If the structure or stratification represented by a particular PC is robust and significant then the two independent marker panels should produce correlated individual PCs. However, the use of a single combination of two non-syntenic SNP panels (e.g. markers on odd vs. even chromosomes[34]) presents an increasing risk of type I or II errors as the number of markers and their overall informativeness diminish, making the PC trend less obvious. Such is often the case in analyses of European population samples, which are less differentiated than the worldwide samples where the test was previously applied[34]. Therefore, the original test was extended to 100 permutations of non-syntenic SNP panels. Under the null-hypothesis of no stratification the P-values are expected to be uniformly distributed on the interval [0, 1]. I calculated Fisher's combined P-value with the program and formula from Dr. Zaykin[82] using 100 Spearman P-values—corresponding to 100 permutated pairs of non-syntenic SNP panels. This is the SKT P-value reported in Table 2.2. It is important to note that PC non-significance can either reflect absence of structure or inability of the marker panel to detect structure.

These two methods (SKT and ANOVA test) were evaluated using simulated groups of individuals with 4 grandparents coming from the same populations (individual genotypes are created based on allele frequencies in the parental populations, with the simsample+.pl Perl script, available at http://www.anthro.psu.edu/biolab/P.upload.zip). Simulations have been performed on artificial populations of 10 to 30 individuals with 4 grandparents coming from the same populations. When two artificial populations are considered only the first PC is significant, when three different artificial populations are considered only the first 2 PCs are significant. This indicates that, when the

informativeness of markers is sufficiently high to see clear stratification, the number of significant PC axes with the SKT or ANOVA test is a good indication of the number of differentiable parental populations. For instance Figure 2.5 represents a simulation based on 3 populations, with 8 individuals per sample; Pop1 has European American allele frequencies, Pop3 has West African allele frequencies, Pop2 is composed of equally admixed individuals. Both the ANOVA test and SKT are highly significant (and ANOVA correlations are very high) only for the first 2 PCs. The decay of percentage of variance indicate that PCs beyond the first two reflect minor patterns, that is the within sample variability which is minor in such stratified data. Finally the fact that the ANOVA test and SKT are in full agreement confirms the proper population labeling of individuals.

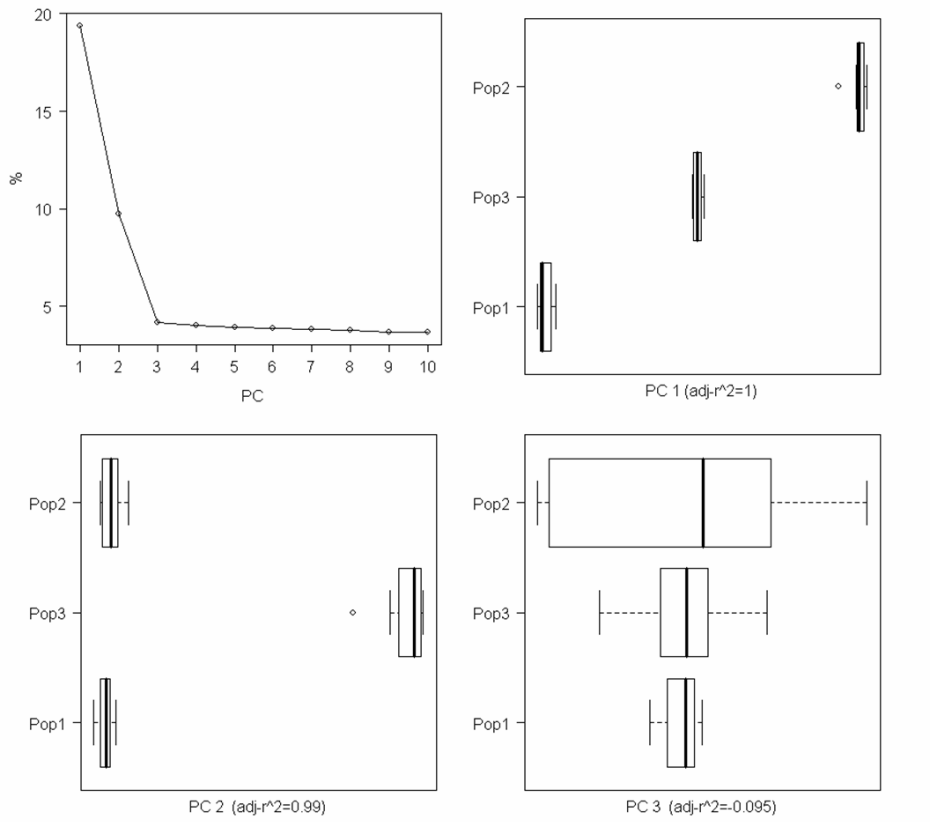


Figure 2.5 – PCoA of three artificial populations of eight individuals each.

Bayesian analysis

Bayesian analysis was performed directly on individual SNP genotypes with the program *structure* (version 2.1)[43,44]. Each individual is represented by a vertical bar, with colors indicating possible admixture proportions from any of the K components, using the companion program *distruct*[83]. Each dataset was run through *structure* at least twice with 40,000 burn-ins and 80,000 subsequent iterations. These parameters were established conservatively by looking for the point at

which several runs would have a stable α, i.e. one which would not vary in amplitude by more than 0.2 (*structure* documentation). The linkage model (using SNP positions) was initially performed on a few datasets but showed little difference from the simpler admixture model. Therefore runs reported here were performed under the admixture model. The non-admixture model was not considered, because the goal is not population assignment of individuals but rather the exploration of population structure and admixture levels, given the knowledge of Europe's long history of gene flow. Also because of historic and geographic proximity European populations are more susceptible to correlated allele frequencies than other world populations, therefore I used the correlated allele frequencies model in *structure*[4]. Posterior probabilities (PP) were calculated using the Bayesian formula (e.g. *structure* documentation). If a PP is not maximum for a given run this doesn't necessarily mean there is no differentiable clusters but may reflect low inter-group diversity of Europe. The clusteredness (G) of each run was calculated using the formula and definition of Rosenberg et al.[4] as "the extent to which individuals were estimated to belong to a single cluster rather than to a combination of clusters". PP and G each offer alternative ways to consider the importance of observed population structure.

Phenotypic measures

Hair and skin were measured using a reflectance spectrophotometer (the DermaSpectrophotometer, from Cortex Technologies, Denmark) to measure the M (melanin)-index. For each participant, three measures of hair pigment were taken in different parts of the head and averaged. Typically the hair M-index was measured only when the subject had enough hair and no dye, which resulted in fewer measurements than for skin. Three M-index measurements were taken under each participant's upper arm, which were averaged into a single value. Eye color was assessed in two different ways. The participants self rated their eye color based on a set of color descriptions (brown, hazel, etc..) before their eyes were photographed with a digital camera—in a cardboard device controlling the amount of light. In a second step these photos were used by T. Frudakis (DNAPrint) to measure the luminance of the iris, based on a simple formula using luminescence, red, blue and green reflectance (T. Frudakis, personal communication).

Data management

I created a database in Microsoft Access and coordinated the keying of the data from the participants forms. European forms were keyed by local helpers, and English forms were entered by Penn State students. The ethnic information was streamlined into labels both at the national and regional levels, accounting for admixture at these multiple levels, most often based on the information

from the 4 grandparents when present. Finally, I cleaned up the database, detecting and fixing outlier values (keying or measurement errors) based on SD between the 3 measures of hair M index and, for skin, left v. right arm mean and SD differences.

Statistical Analysis

The Spearman rank correlation coefficient, ρ (rho), a non-parametric version of the Pearson correlation, was used to measure correlation between pigmentation measures and ancestry measures, in order to avoid making assumptions about the frequency distribution of the variables.

The standard Wilcoxon test was used to measure the significance of differences between the distributions of test results from different population samples. The Wilcoxon test, also known as the Mann-Whitney test, can be considered as a non-parametric T-test in the sense that it makes no assumption in the shape of the distributions being compared. When multiple tests were performed, such as between all pairs of samples, a Bonferroni correction was applied.

Computations were performed and most figures were generated mainly using two high-level scripting languages; the *bash* Linux shell and DOS batch under Windows XP. A library of Perl scripts was created to execute the varied data processing and calculation. The R statistical package[84] was used to write scripts for statistical analyses and graphic representation of data and results. The main scripts are available at http://www.anthro.psu.edu/biolab/P.upload.zip. See scripts with extensions .sh, .pbs (Unix shell scripts) and .bat (Windows batch scripts) for example of how Perl and R scripts can be used.

Results

DNAPrint Ancestry Tests

All available individuals typed for 313 markers of EuroDNA 1.0 were selected as described above (Eliminating Outliers and Relatives). The worldwide AncestryByDNA 2.5 test (Figure 2.6A) already shows major differences between average profiles of South Asians (59%EU, 27%EA, 8%NA), East Africans (45% EU, 48% AF, 5%NA), Arabs (79%EU, 13%AF), Turkish (85% EU, 7% EA, 5% NA), Ashkenazi Jewish (87%EU, 5% EA), Greeks (90%EU, 5%NA), western Europeans (91-94%EU, other components <4% on average).

The MLE ancestry components based on the complete panel of 313 AIMs from EuroDNA 1.0 are shown for available individuals in Figure 2.6B. The test clearly distinguishes the South Asians (averaging 83%SA on our sample, nearly 100% for the parental individuals). Arabs and Iranians whose genotypes were used for the ME component are the ones with nearly 100% ME in Figure 2.6B whereas other Arabs and Iranians do not (rarely over 50% ME). A similar situation prevails for the SE component on Figure 2.6B, where Turkish and Greek parental individuals are the ones with 100% SE, whereas other Turks and Greeks have inconclusive levels—and so do other southern Europeans like Iberians, Italians and southern French. Finally, NE parental individuals have between 75% and 100% of this component; however other northern Europeans from Iceland, Germany, Ireland and England "ranged between 55% and 94% of SE, frequently having other components higher than 20%.

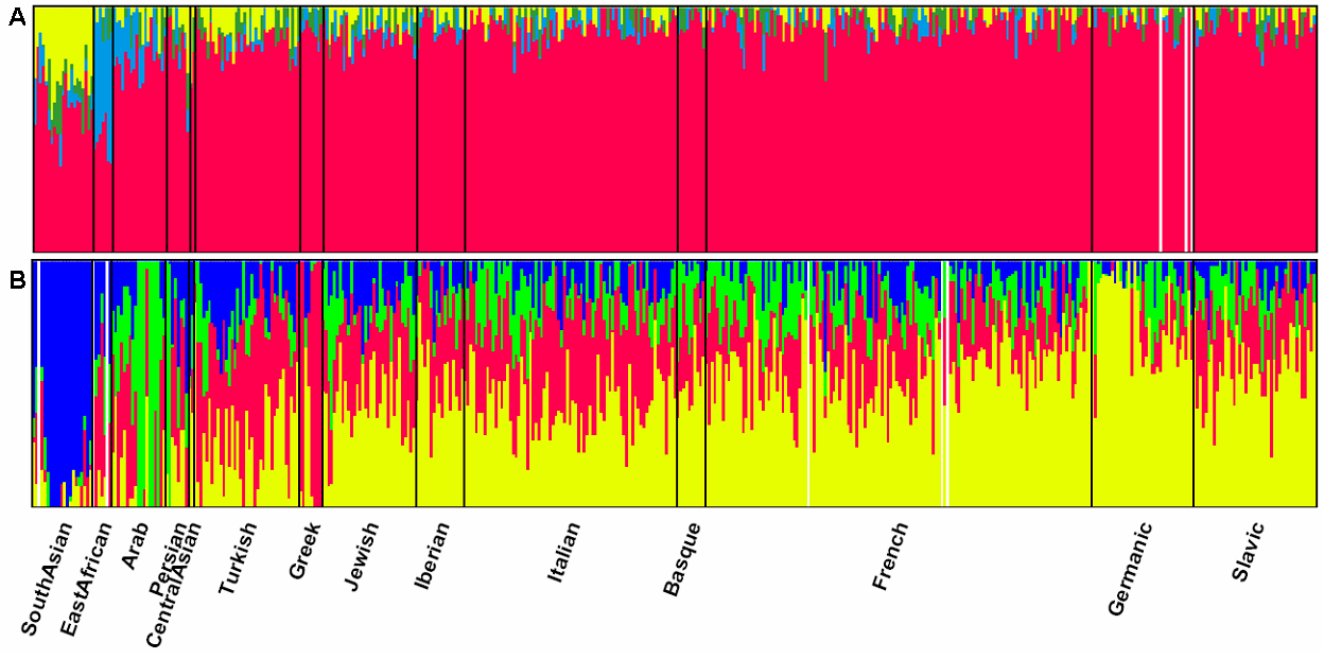


Figure 2.6 – DNAPrint ancestry test results. Each stripe represents an individual. White stripes are missing results. ***A,*** AncestryByDNA 2.5 (Red=EU, Blue=AF, Yellow=EA, Green=NA) and ***B,*** EuroDNA 1.0 (Yellow=NE, Red=SE, Green=ME, Blue=SA)

Raw Genotypes Analyses with *structure*

Using a selection of all available individuals typed for EuroDNA 1.0's 313 markers (see *Eliminating Outliers and Relatives* in *Methods*), I ran the program *structure* for a range of putative clusters from K = 1 to 8 (Figure 2.7).

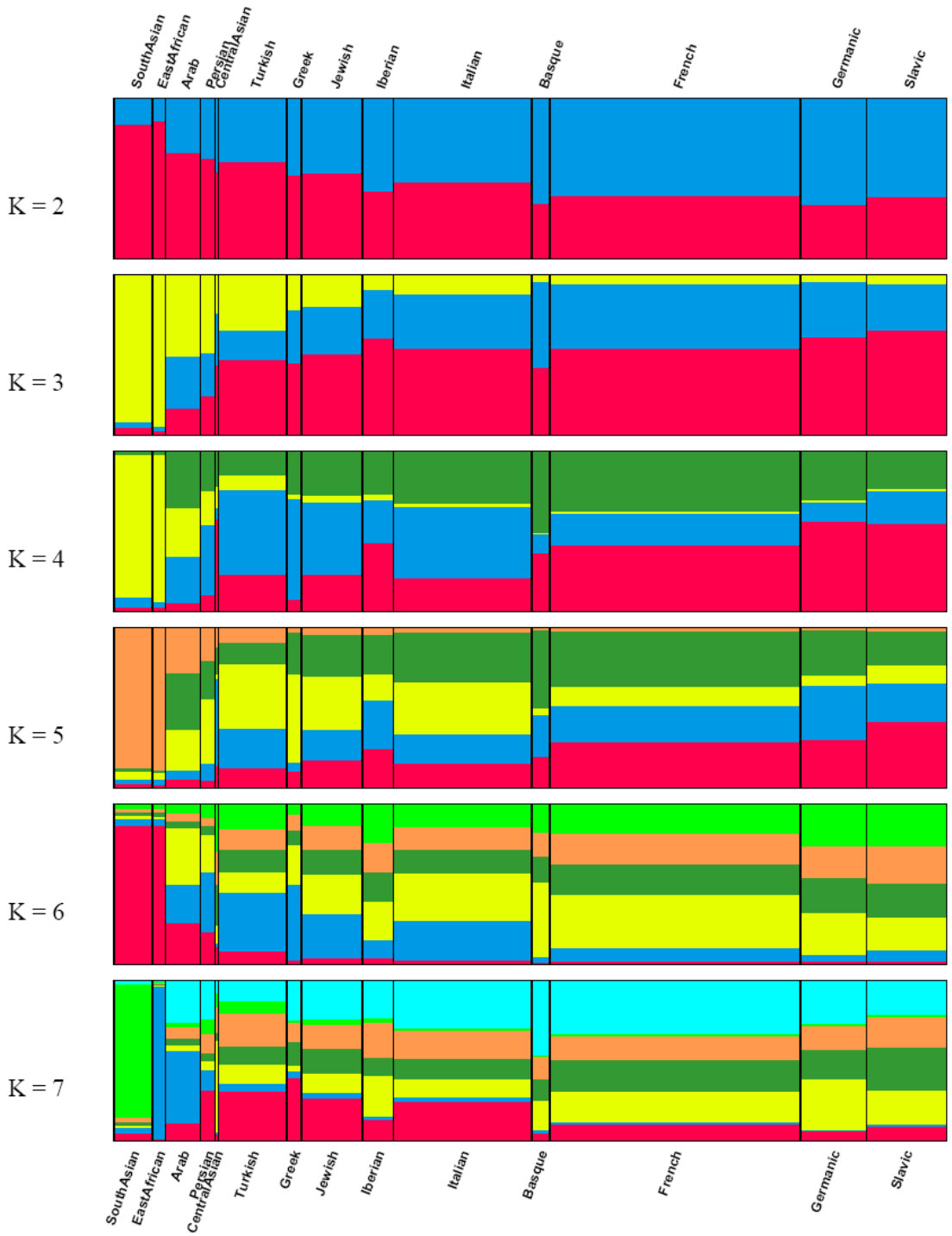


Figure 2.7 – Results from *structure* with K = 2 to 7 clusters, using the full set of individuals. The cluster percentages are averaged for each population sample. The width of stripes is proportional to sample size. "Germanic" includes speakers of Germanic languages (mostly Irish, Icelandic, and some Coriell Northern Europeans). "Slavic" includes mostly Russian and Polish individuals.

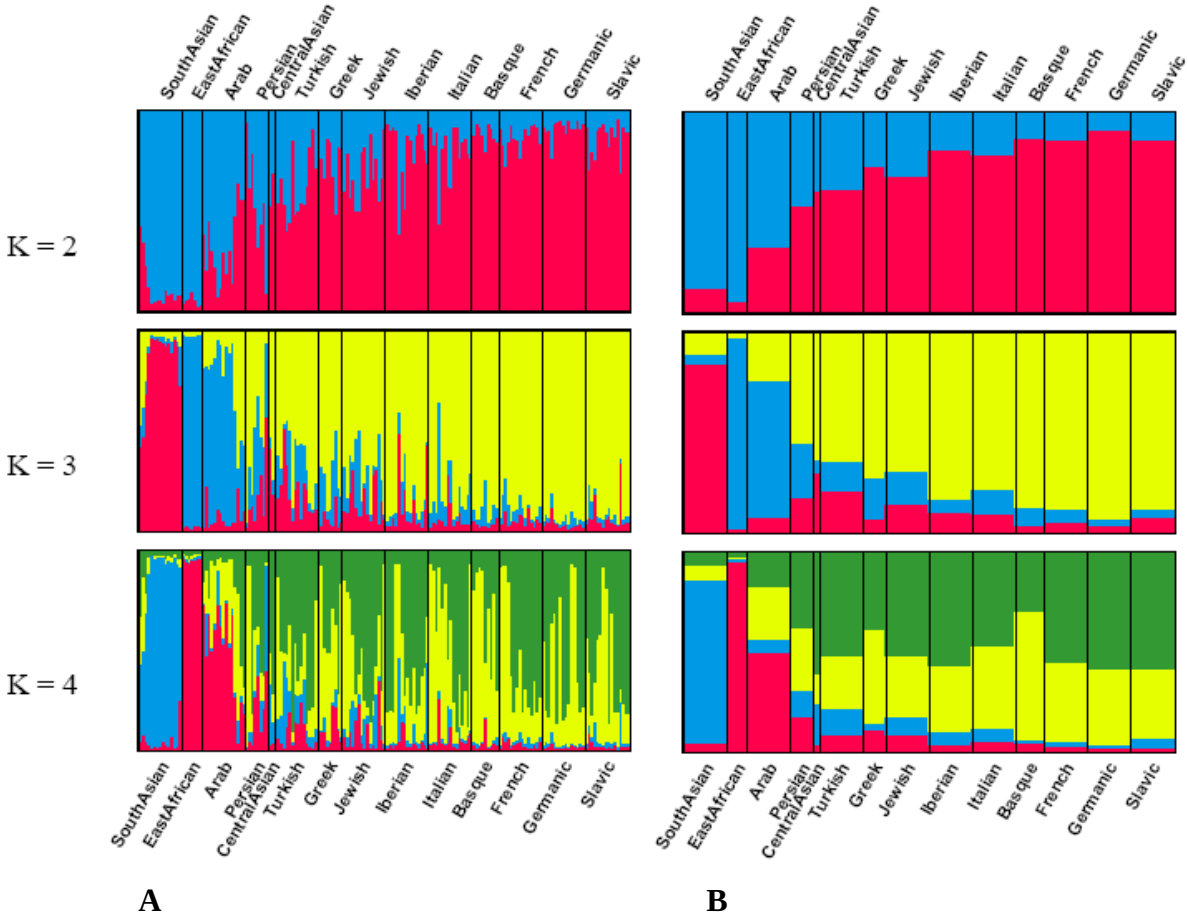
The posterior probability (PP) and the overall clusteredness (G), respectively ~100% and 75%, are the highest for K=4 (Figure 2.7). In this model the South Asians and East Africans define the most distinct clusters (respectively G=86% and G=90%). Two other clusters approximately reflects a north-south axis, better observed in the next favored K=3 model (G=73%). Finally the fourth cluster, absent from South Asians and East Africans, is somewhat higher among the Basque (51%). These last three components are not clearly explained by information on individuals' detailed geographical location.

In order to check for possible spurious patterns created by the unbalanced sample sizes, I selected at most 15 random individuals picked from each sample larger than 15 individuals. The *structure* analysis with this reduced set displays a clearer picture (Figure 2.8). Now the favored model (PP ~100%) is unambiguously K=3, with by far the highest clusteredness (G=72%). From left to right, the clusters of Figure 2.8 (A and B) are defined as follows:

1 – The first cluster includes most South Asians (except the two Pakistani who have <50%). The five individuals who are also used as SA parental individuals in EuroDNA 1.0 have consistently over 92% of this cluster. High levels of this cluster are also found in one Central Asian (Afghan, 57%) and one Turkish (52%). The one Italian who has over 50% of this cluster is of mixed Sardinian and Sicilian ancestry. Other individuals with >10% of this cluster appear in several European samples without apparent justification.

2 - The next cluster is defined by East Africans and Arabs (from Tunisia, Morocco, Saudi Arabia and Palestine), except the last three Arabs who have less than 40% (two Lebanese and one Syrian). Other individuals who come close to 50% of this component are found among the Iranian, Turkish, Jewish and Greek samples. One Italian, from Sicily, reaches 58%.

3 - Finally the vast majority of geographically-defined European individuals, as well as most Jewish and Turkish, have the highest affinity to this third cluster. Within these samples, the percentage of this cluster approximately decreases from the most northern to the most southern groups, a trend which is the inverse of the first two clusters. The next groups to share some of these components are the Arabs (23%) and South Asians (10%) whereas it is negligible in East Africans.



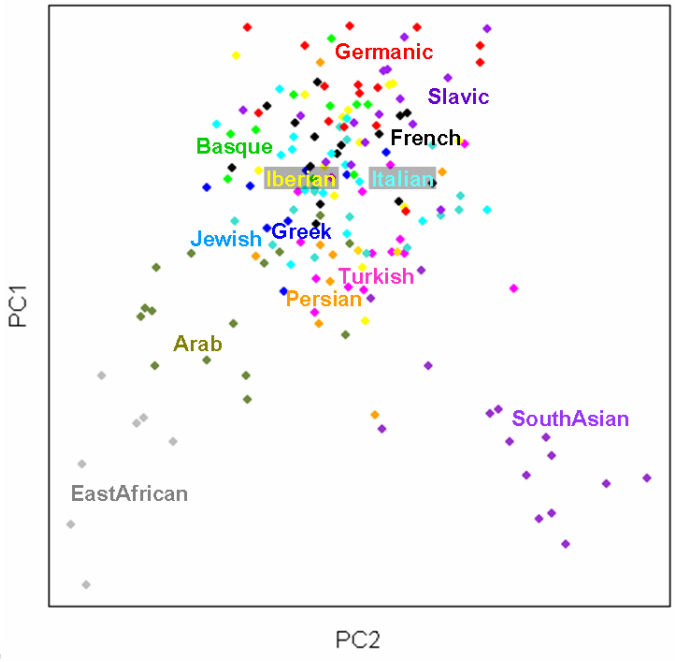


Figure 2.8 – Results from *structure* using the reduced set of individuals. ***A,*** each stripe representing an individual, and ***B,*** with cluster percentages averaged for each population, ***C,*** PCoA.

Raw Genotypes Analyses with PCoA

For consistency with the *structure* analysis, PCoA and significance tests were performed on both the full and the reduced set of individuals (as defined above). In both sets the percentage of the variance explained by each PC drops quickly and stabilizes after the 3rd PC (Figure 2.9), indicating that patterns shown by PCs >3 are either less important or less well detected by the AIM panel.

For the full set, the percentage of the variance explained by all top 10 PCs is about two times lower (Figure 2.9), which confirms that, similar to *structure*, having too many individuals of similar genetic ancestry decreases the clustering power of PCoA. This is also consistent with the fact that ANOVA correlations and the significance of the top PCs are higher in the reduced set (Table 2.2). For the full set, the first 3 PCs are highly correlated with population labels (ANOVA test, Table 2.2). PC4 and PC5 are significant with both tests (Table 2.2, Full Set) and both mark the distinctiveness of the two Central Asians (not shown) but the correlation is small, probably due to the small sample size of this group. PC7 (ANOVA P<0.001) is a reminiscence of the separation of the East Africans (not shown) also seen in PC2 and PC3.

Table 2.2 – PCoA significance using SKT and ANOVA test.

	Full Set		Reduced Set	
PC	**Adj. R2** (P) †	SKT P ‡	**Adj. R2** (P) †	SKT P ‡
1	**0.66** (<0.001)	<0.001	**0.77** (<0.001)	<0.001
2	**0.10** (<0.001)	<0.001	**0.60** (<0.001)	<0.001
3	**0.33** (<0.001)	NS	**0.12** (<0.005)	<0.001
4	<0.1 (<0.05)	<0.001	**0.12** (<0.005)	NS
5	<0.1 (<0.001)	<0.001	<0.1 (NS)	NS
6	<0.1 (NS)	NS	<0.1 (NS)	NS
7	**0.11** (<0.001)	NS	<0.1 (NS)	NS
8	<0.1 (<0.05)	NS	<0.1 (NS)	NS
9	<0.1 (NS)	NS	<0.1 (NS)	NS
10	<0.1 (NS)	NS	<0.1 (<0.05)	NS

† ANOVA correlation coefficient (adjusted R-squared) between PCs and group labels in bold when >0.1, with corresponding P-value in parenthesis. NS means not significant at the 0.05 level. The non-parametric Kruskall-Wallis test did not disagree with the ANOVA tests, and the Bartlett test of non-equality of variances among group labels was never highly significant.

‡ Combined P-value calculated for SKT as described in Material and Methods.

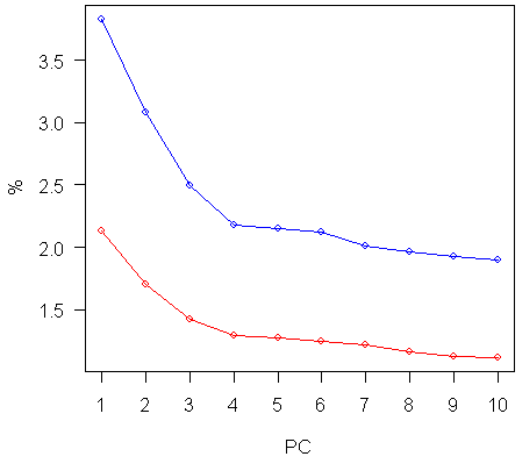


Figure 2.9 – Decay of percentage of the variance explained by each PC for the full set (red) and the reduced set (blue).

In PCoA of the reduced set (Figure 2.8C and Figure 2.10) PC1 approximately follows a north-south axis; the East Africans, South Asians stand out as most distinctive from Europeans and Middle Easterners. PC2 separates the East Africans and South Asians (Figure 2.8C and Figure 2.10). PC1 and PC2 together are consistent with geography and my 10k microarray analysis of European individuals[85]. Similarly to the *structure* analysis the Arabs show a marked shift toward the East Africans. Despite large overlaps, PC3 is driven by a slight separation of the Basques from other Europeans, and more differentiation among the East African, South Asian and Arab samples. PC4 is reminiscent of the north-south axis in the European samples, and Central Asians show a slight distinctiveness. However the Central Asian pattern is not interpretable at this point because only two individuals fall under that label, which again may be the reason why the PCs where they appear most distinctive (not shown) have low significance.

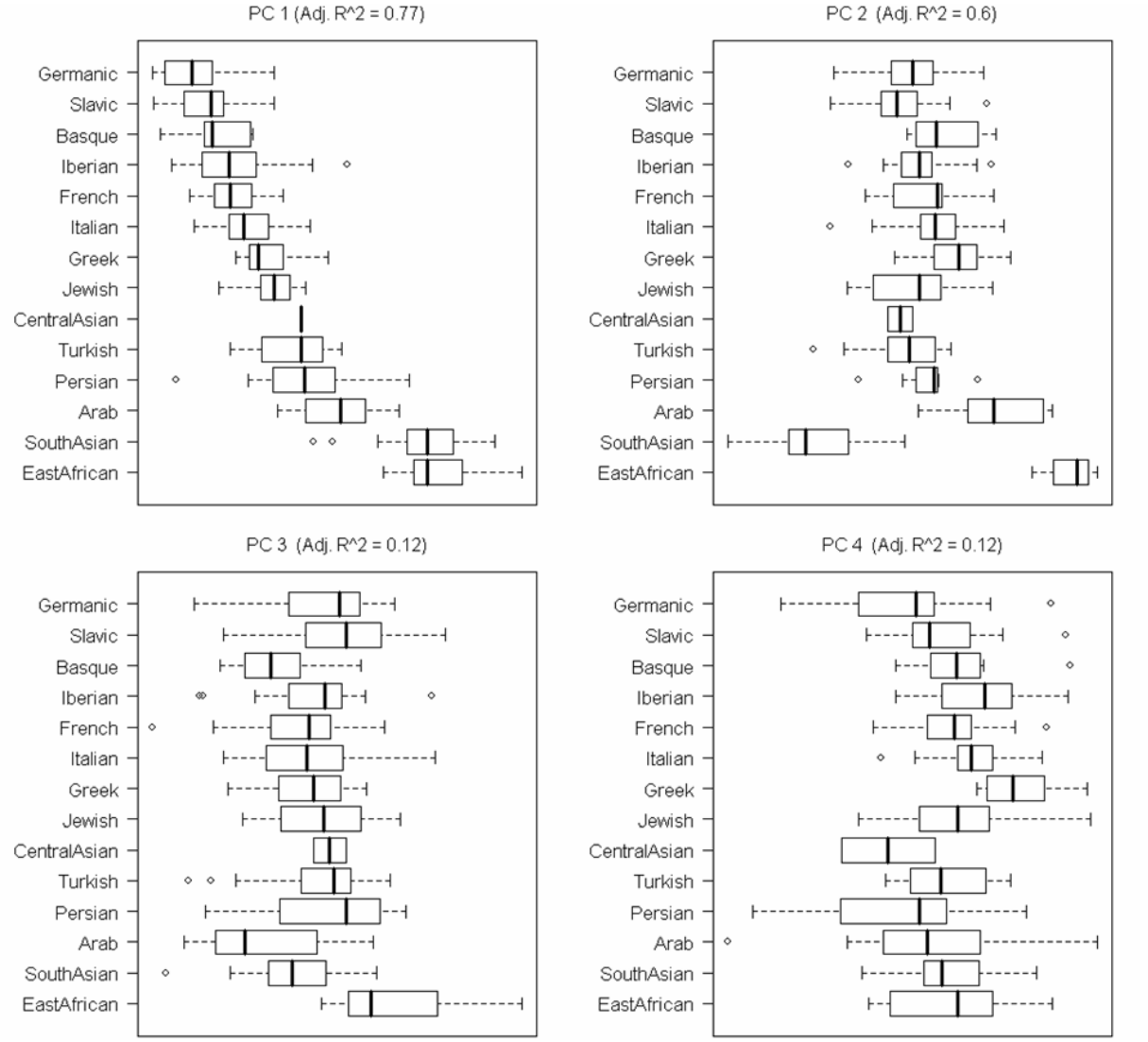


Figure 2.10 – Top four PCs (reduced set of individuals), with ANOVA correlation coefficients from Table 2.2.

Phenotypes

Hair Pigmentation

On a global level the distribution of hair pigmentation appears mostly normal (Figure 2.11A), with a slight skew toward the left, i.e. lighter pigmentation (median=117.2, mean= 116, SD=24) which is not significantly different from the European-derived distribution (Figure 2.11B). Some population samples presented a rather different picture (Figure 2.12 and Table 2.3). The Irish sample peaks around M~100, whereas the Portuguese and Italian both have a single mode at M~130. The Polish, French and Jewish Americans have a bimodal distribution reflecting previously observed peaks at ~100 and ~130.

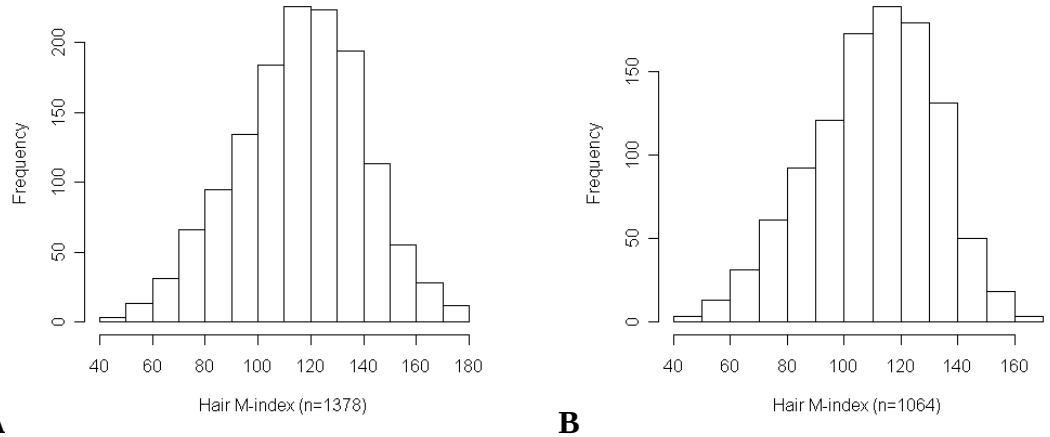


Figure 2.11 – Hair pigmentation distribution for ***A,*** all available individuals, and ***B,*** European-derived individuals only

Table 2.3 – Differences between hair pigmentation distributions

	French	Irish	Italian	Jewish	Polish
Irish	<0.05	-	-	-	-
Italian	<0.1	**<1e-9**	-	-	-
Jewish	NS	<0.0005	NS	-	-
Polish	NS	NS	**<1e-9**	<0.01	-
Portuguese	NS	**<1e-9**	NS	NS	**<1e-9**

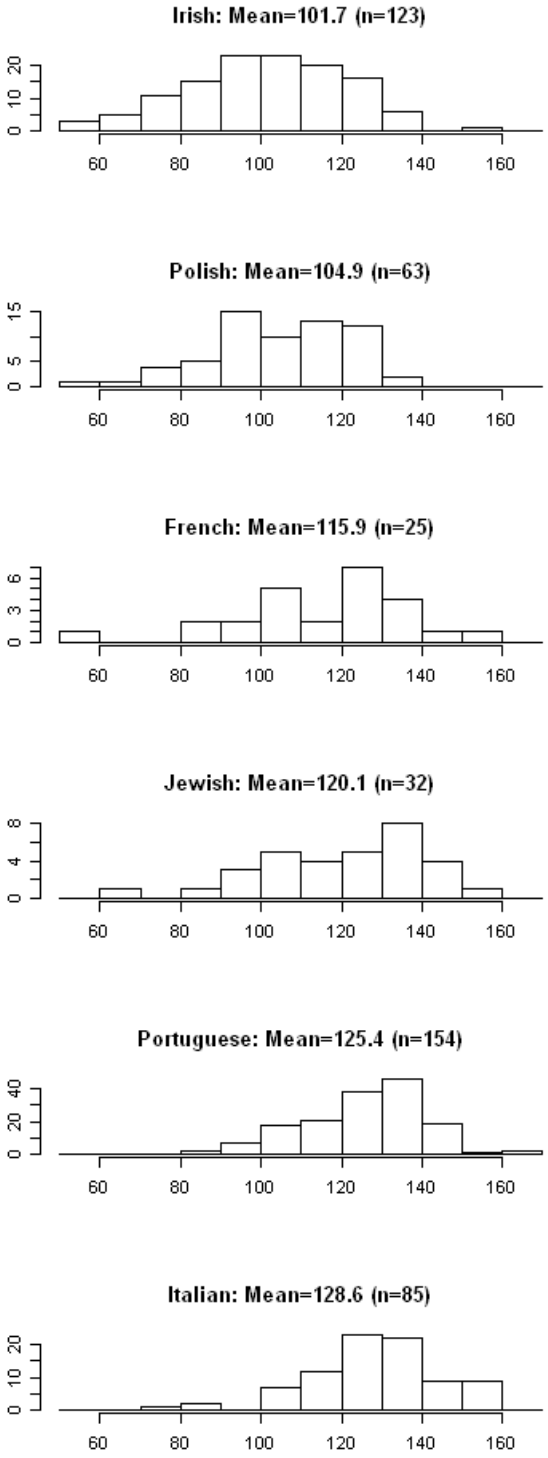


Figure 2.12 – Hair pigmentation (M-index) distribution for largest European samples

Skin Pigmentation

Skin pigmentation for all individuals spans a range of M-index between 20.7 and 74.8 (mean=32.2, median=30.5, SD=7.2) and is highly skewed to the left in this dataset (Figure 2.13A) because it contains a majority of individuals of European ancestry. Figure 2.14 and Table 2.4 show that, despite important overlap among all samples, the Irish are significantly lighter than all other population samples (mean M = 26.4). The Wilcoxon difference between the Irish and the nearest sample (Polish) is -3.3 (95%CI = -3.9 to -2.7). The Polish, Portuguese, Jewish, French and Italian samples have very similar means (Figure 2.14), not significantly different from one to the next (Table 2.4). Only the Italian sample appears significantly darker than the Polish and Portuguese samples (Table 2.4; Italian column).

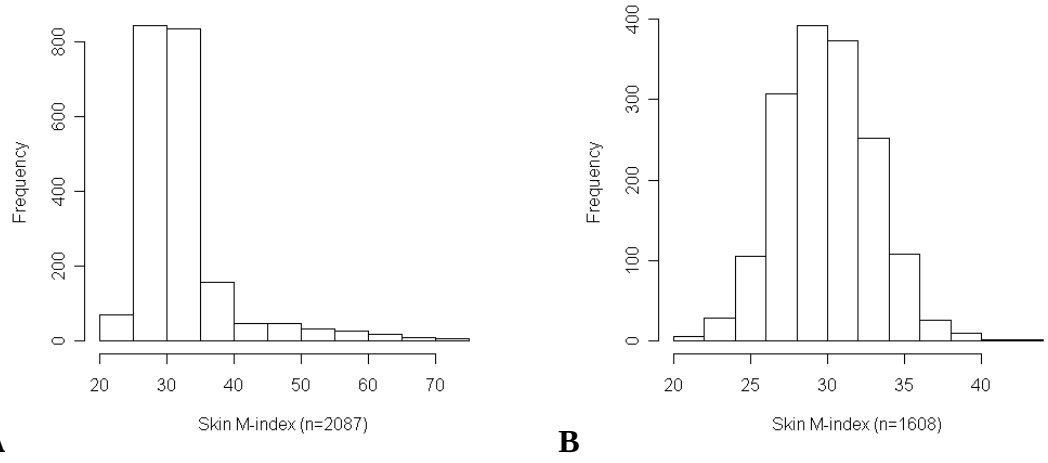


Figure 2.13 – Skin pigmentation (M-index) distribution for ***A,*** all available individuals, and ***B,*** European-derived individuals only

Table 2.4 – Differences between skin pigmentation distributions

	French	Irish	Italian	Jewish	Polish
Irish	**<1e-9**	-	-	-	-
Italian	NS	**<1e-9**	-	-	-
Jewish	NS	**<1e-9**	NS	-	-
Polish	NS	**<1e-9**	<0.005	NS	-
Portuguese	NS	**<1e-9**	<0.005	NS	NS

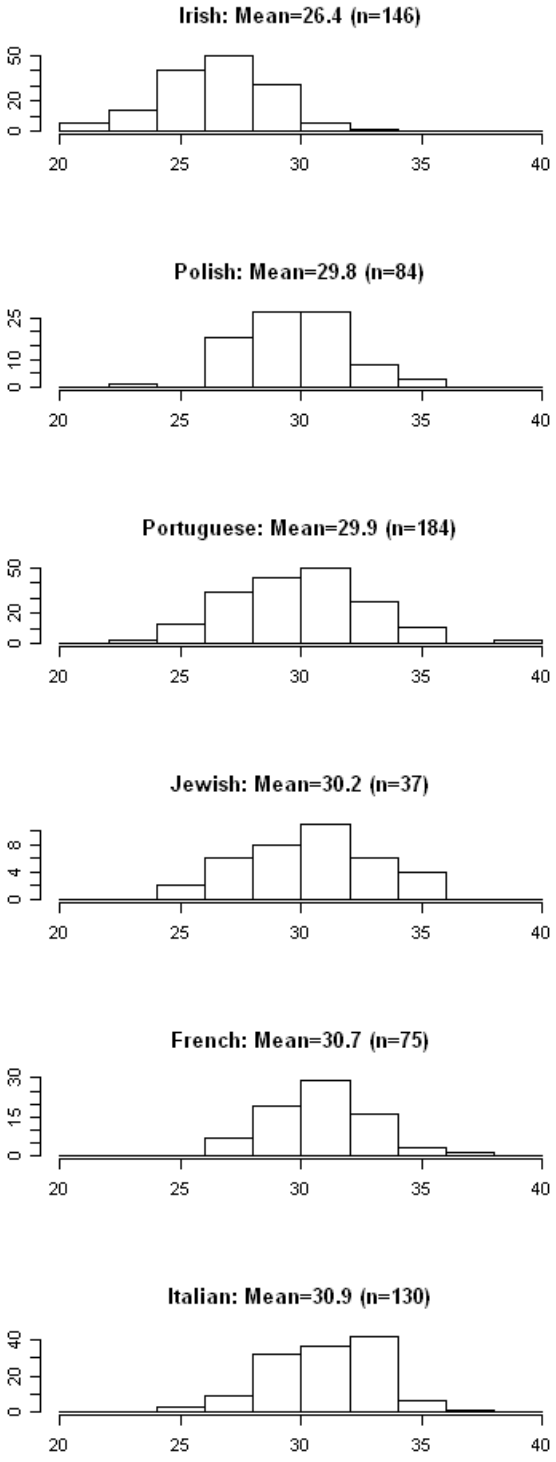


Figure 2.14 – Skin M-index distribution for Europeans and Jewish American samples.

Eye Pigmentation

The distribution of iris luminance is bimodal among Europeans (Figure 2.15), mainly reflecting light eyes v. dark eyes. Figure 2.16 shows individual population sample distributions and Table 2.5 the corresponding significance of differences. With one peak around 3 and a large tail to the lower luminance values, the Irish have significantly lighter eyes than all other population samples. The Polish display an approximately bimodal distribution with a large fraction of intermediate values. The French have mostly darker eyes but the distribution has a heavy right tail reflecting a large proportion of light and intermediate eye colors. The Italian, Portuguese and Jewish Americans have a similar distribution with one peak around 1.5 luminance and very few individuals with light eyes. As a result, Europeans and European Americans (Figure 2.15B) display a bimodal distribution peaking at ~1.5 (dark eyes) and ~3 (light eyes), which is also true when plotted separately (not shown).

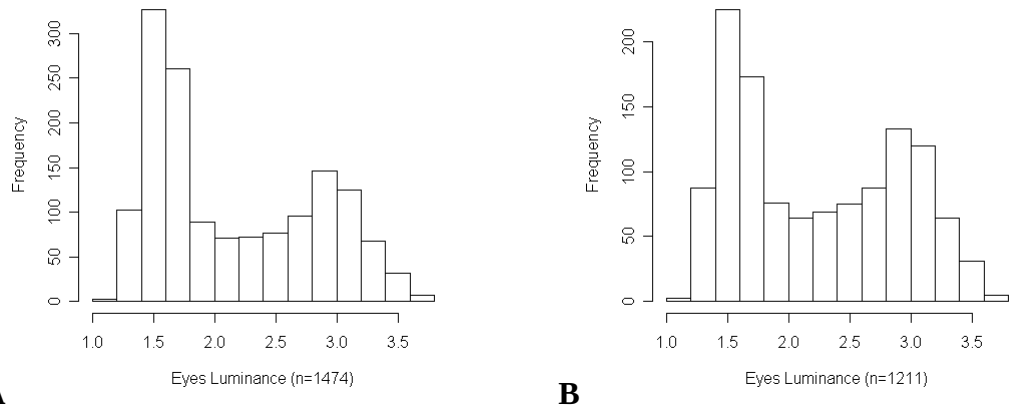


Figure 2.15 – Eye luminance for ***A,*** all individuals and ***B,*** European-derived individuals

Table 2.5 – Differences between eye luminance distributions

	French	Irish	Italian	Jewish	Polish
Irish	<1e-3	-	-	-	-
Italian	**<1e-9**	**<1e-9**	-	-	-
Jewish	<0.01	<1e-3	1	-	-
Polish	1	<0.01	**<1e-9**	<0.05	-
Portuguese	<1e-9	**<1e-9**	1	1	**<1e-9**

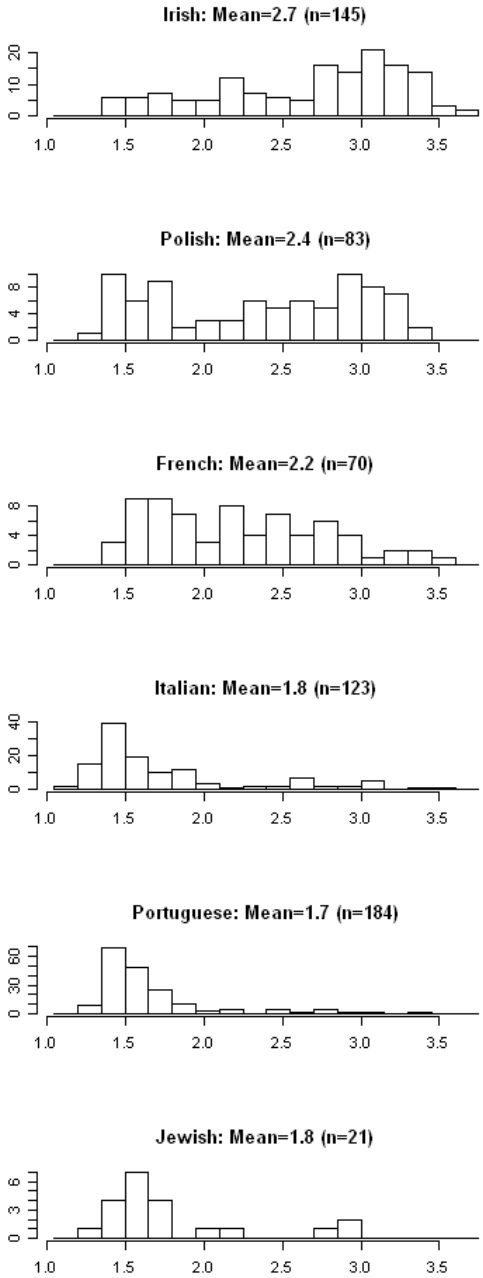


Figure 2.16 – Eye luminance distribution for largest European samples and Jewish Americans.

Phenotype-Ancestry Correlation

Hair Pigmentation

The AncestryByDNA 2.5 results all show significant correlation with hair M-index (Figure 2.17). For all ancestry components the trends are mainly driven by the European vs. non-European hair color differences since Europeans have the lowest hair M-index of the world populations considered here. Not surprisingly, hair pigmentation increases as Western European (EU) decreases, and conversely the three non-European components are positively correlated with hair pigmentation.

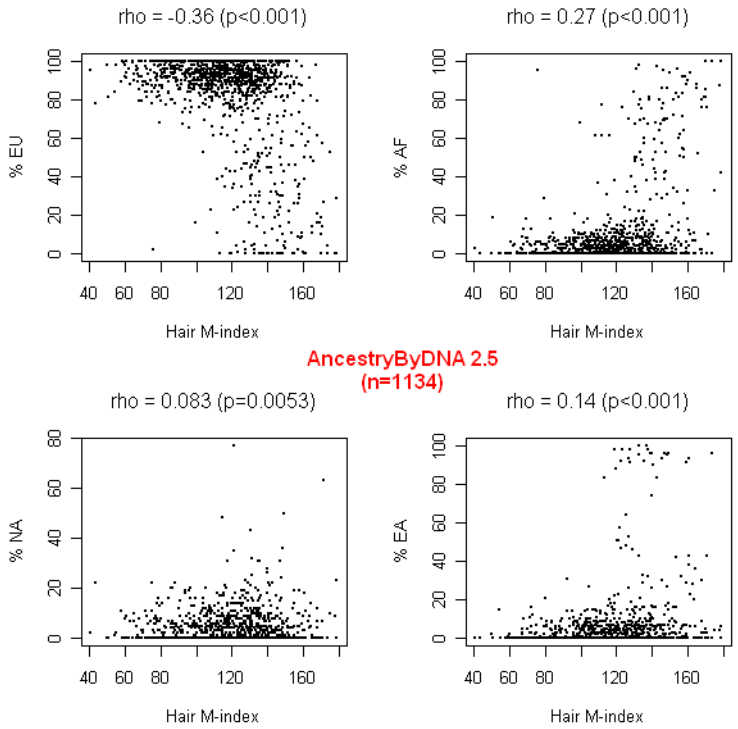


Figure 2.17 – Spearman correlation and plotting of hair M-index with genetic ancestry for all individuals with hair data (Table 2.1)

The ME component of EuroDNA 1.0 does not correlate with hair pigmentation, but NE and SA components do, as well as SE to a lesser extent (Figure 2.18). The correlation coefficient for SA is mainly driven by the South Asian individuals. NE is negatively correlated with hair pigmentation whereas SE is positively correlated.

Finally, if we only consider geographically-defined European populations, only NE and SE correlate significantly with hair pigmentation (Figure 2.19).

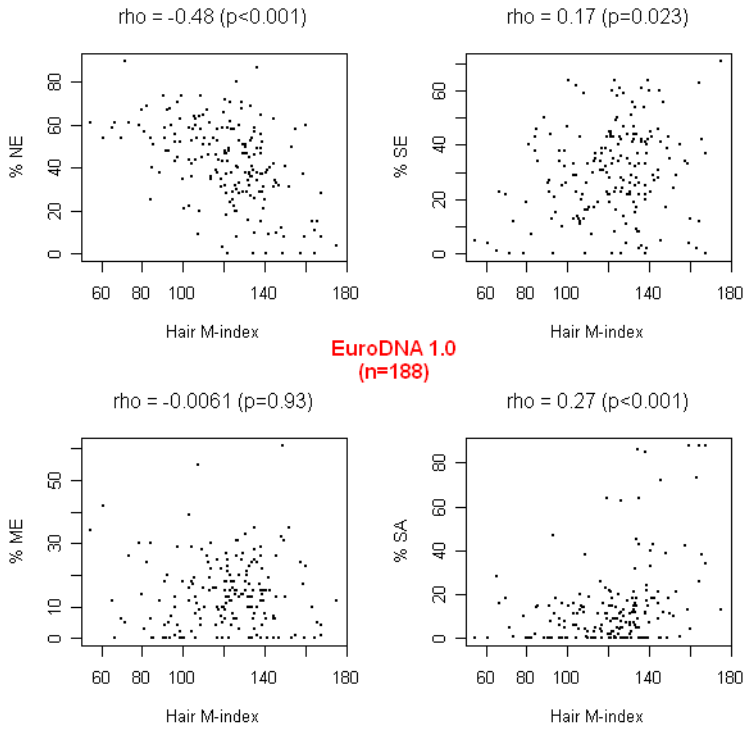


Figure 2.18 – Spearman correlation and plotting of hair M-index with genetic ancestry for all European individuals with hair EuroDNA 1.0 results and hair data (Table 2.1).

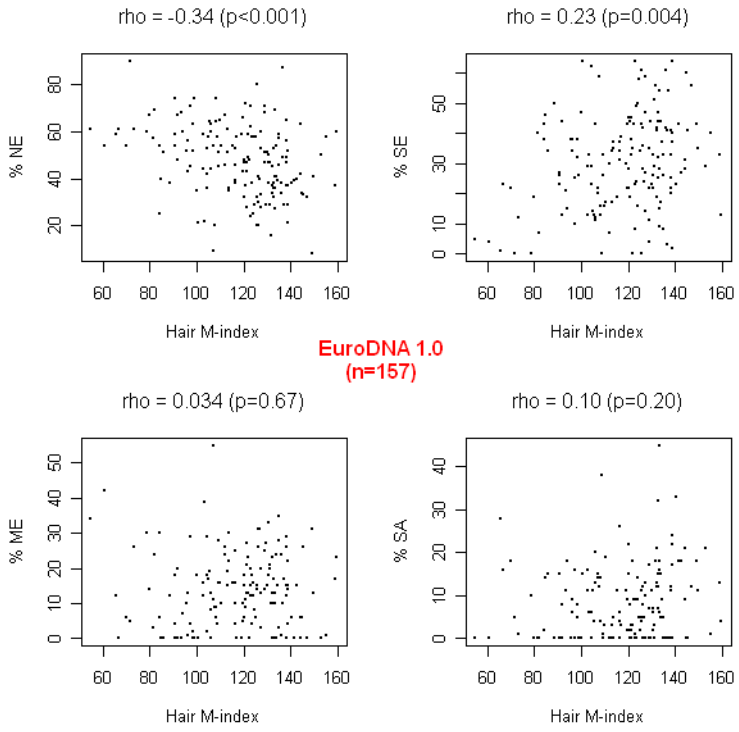
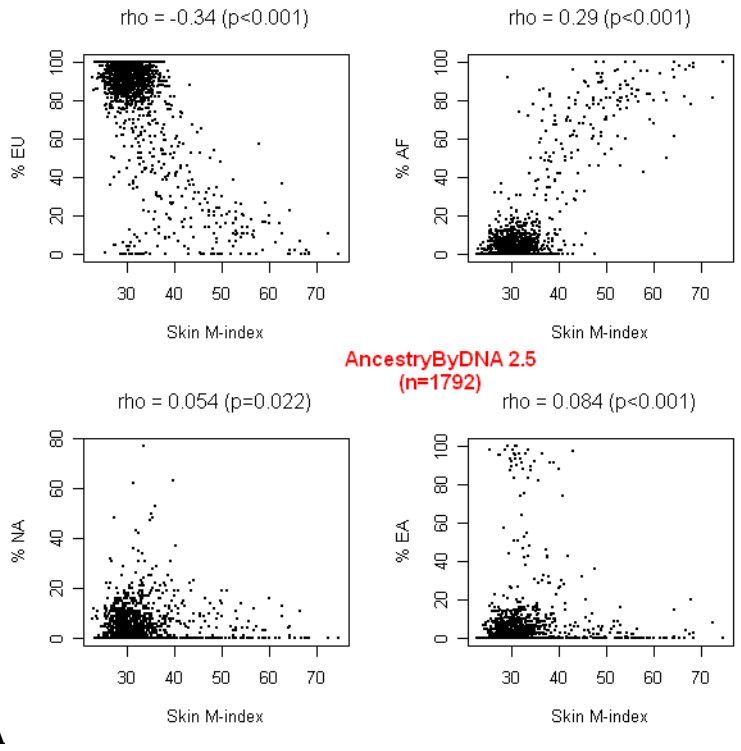


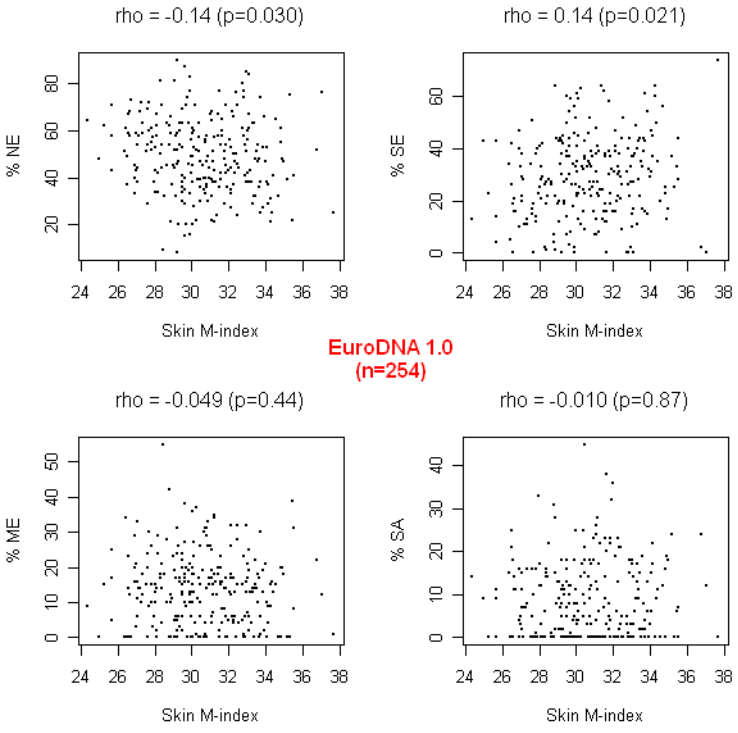
Figure 2.19 – Spearman correlation and plotting of hair M-index with genetic ancestry, keeping only individuals from traditional populations of Europe with hair data (subset from Table 2.1).

Skin Pigmentation

All four components of the AncestryByDNA 2.5 test (EU, AF, NA and EA) show significant correlations with skin M-index (Figure 2.20A). The trends are mostly driven by the European vs. non-European skin color differences, since Europeans have the lowest skin M-index of the world populations considered here. The most important correlations are along the European-African axis, which is not surprising considering the presence of many admixed individuals from these two origins. With EuroDNA 1.0 only NE and SE are correlated with skin pigmentation, though only mildly ($P<0.1$), with equal absolute correlation 0.14 (Figure 2.20B below).



A

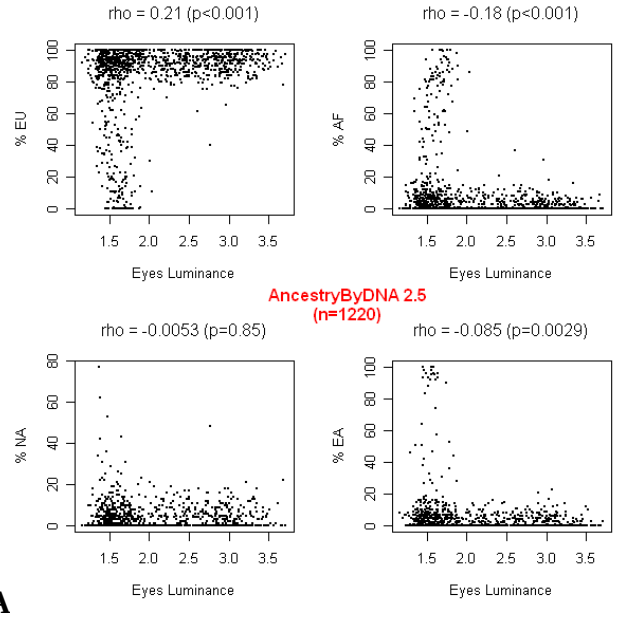


B

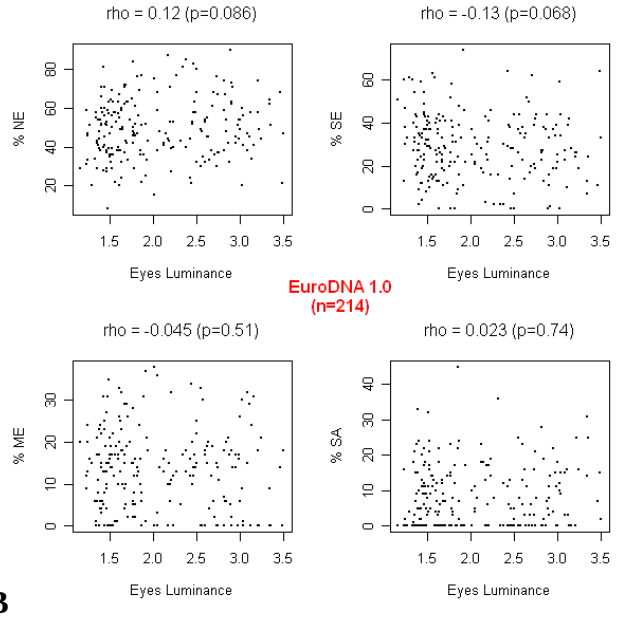
Figure 2.20 – Spearman correlation and plotting of skin M-index ***A,*** with worldwide genetic ancestry using world populations and ***B,*** with European ancestry and individuals.

Eyes Pigmentation

Significant correlations between eye pigmentation and AncestryByDNA 2.5 components EU, AF, and EA are mainly driven by the wide range of eye luminance in Europeans (Figure 2.21A). Correlations with EuroDNA 1.0 are not significant for ME and SA (even when Middle Easterners and South Asians are included, not shown) and only mildly significant for NE and SE (Figure 2.21B). There is a slight positive correlation between NE and eye luminance and a slight negative correlation with SE.



A



B

Figure 2.21 – Spearman correlation and plotting of Eye luminance ***A,*** with worldwide genetic ancestry using world populations and ***B,*** with European ancestry and individuals.

Discussion

Archeological evidence shows that agriculture and Indo-European languages spread from modern-day Pakistan to the east of the Indus Valley during the Harappan phase (between 4,500 and 4,000 years ago) and then diffused into central India. Other forms of agriculture developed in the northeast of India (8,000–5,000 years ago) and spread west and south from there. Several regional farming populations may also have been in place prior to the immigration of Indo-European speaking farmers[25]. Somewhere between the demic and cultural diffusion models, we may reasonably accept that some gene flow occurred when Neolithic farming cultures spread from the Near East, in both directions, west toward Europe and east into South Asia[14]. Therefore from a theoretical standpoint it is hard to imagine how the SA component could represent a parental population for a typical European individual.

In practice, the SA component is well defined by the EuroDNA 1.0 markers (Figure 2.6), as reflected also by the distinctiveness of South Asian individuals in direct Bayesian analysis (Figure 2.8A and 2.8B, K=4). The 12 Indians in this dataset have 70% SA on average (SD=17.9). This is less than the ~100% that would be expected, but significantly higher than the European (n=405) mean of 7.7% ($P<0.001$). The 7 Iranians in this dataset have on average 16.6% SA, not significantly more than Europeans ($P=0.11$). The 20 Arabs have 10.5% which is not significantly higher than Europeans ($P=0.46$). Therefore, the presence of a South Asian ancestry component in a European individual is difficult to interpret, other than to say that ancient common ancestry may have caused this to happen. The SA component is also difficult to interpret as admixture for a West Asian individual such as an Iranian or Arab; their higher SA level may simply reflect geographical proximity. Similarly, the one Afghan individual appears as a mix of SA and ME, reflecting the geographical position of Afghanistan between the Middle East and the Indian Peninsula.

In view of the both recent and Neolithic migrations from North Africa and the Near East, the ME component appears as to be a more reasonable parental group for Europeans (and South Asians too). The Middle East is an historical and political region which includes most of the Arab-speaking world (from Southwest Asia to North Africa), Iran and sometimes Pakistan, Afghanistan and Turkey. Both DNAPrint's ancestry test results and the multivariate analysis on the same markers (Figures 2.6, 2.8 and 2.10) show that this wide area presents high levels of genetic diversity, which is little surprising considering its geographical position at the crossroads of three continents. Probably more than one parental population may emerge from this region in the future.

Parental individuals defining the ME component are a composite group of Arabs of partly North African origin and a few Iranians (Appendix B). Therefore even in the unlikely event that such a genetic entity as Middle Eastern exists, the ME parental individuals do not properly represent the expected geopolitical range. Many Europeans do display a significant ME percentage although it was expected to be rather limited in most Europeans because Middle Eastern immigrants are a minority of recent origin. If the idea of a Middle Eastern parental population was meant to represent a contribution from Neolithic farmers, parental samples from the Near East may have been more appropriate. The fact that the 321 Europeans typed in this study have on average 13.3% of ME (SD=10.6) is more likely to reflect inaccuracies inherent to the test rather than real admixture from the Middle East. Additionally, the 5 Arabs in the dataset not used as parental ME individuals only have 29.2% ME on average (SD=20.5), which is not significantly more than the Europeans (P=0.8). This lack of differentiation from Europeans make the interpretation of the ME component even more problematic.

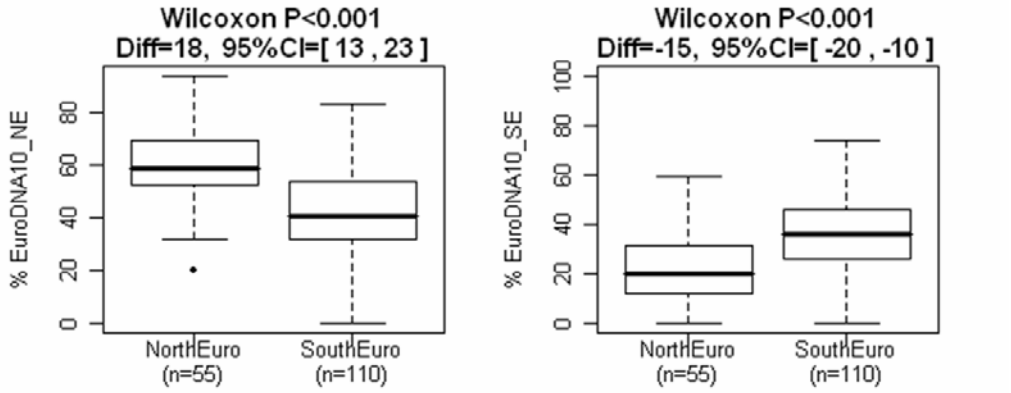


Figure 2.22 – EuroDNA1.0 NE and SE distributions on selected Europeans from the North (Dutch, English, Icelandic, Irish, Russian, Swiss, Polish, Russian, Danish, German, Norwegian, Lithuanian) and South (Armenian, Bulgarian, Greek, Spanish, Italian, Portuguese). Differences for ME and SA are not significant (not shown).

Finally, the Northern European (NE) and Southern European (SE) components are validated by the result that individuals of northern ancestry have significantly higher NE and lower SE component than southern individuals (both Wilcoxon P<0.005). However there is a large amount of overlap (Figure 2.22). This could be interpreted as lack of genetic differentiation between northern and southern Europeans, but other evidence shows that the north-south axis is expected to be more marked in Europe[46,85]. In fact, if we look at hair and eye color distributions in six traditional European populations (Figure 2.14 and 2.15), these phenotypes appear to be better predictors of north-south ancestry than EuroDNA 1.0 (Figure 2.4B). Similarly, skin color difference between the Irish sample and continental samples (Figure 2.12 and Table 2.3) are relatively larger than Irish differentiation

observed with the 313 AIMs of EuroDNA 1.0, if any (Figure 2.6B, 2.6, 2.7). Therefore individual differences in NE and SE levels must be interpreted with great caution.

In comparison with the worldwide AncestryByDNA 2.5 test the likelihood intervals of EuroDNA 1.0 components are more than two times wider on average (respectively 23.8 (n=530) and 10.4 (n=2,088) for the 2X MLE intervals), reflecting the very low resolution power of EuroDNA 1.0.

Conclusion

The EuroDNA 1.0 test is a first generation ancestry test for Europeans and presents important limitations in describing an individual's ancestry. As in the worldwide AncestryByDNA 2.5 test, the model of 4 parental populations offers a convenient presentation, but the 4 components of EuroDNA 1.0 lack descriptive power for the European genetic landscape. The model may be improved by eliminating the SA component but resolution power is unlikely to increase. Despite using a large number of individuals of a wide variety of backgrounds only a very slight north-south pattern of genetic differentiation was detected within Europe (Figure 2.22); using PCoA and Bayesian techniques also showed slight separation with large overlap (Figure 2.8). Individuals from regions around Europe (north and east Africa, south, west and central Asia) define clear north-south and east-west axes, and possibly more (Figure 2.8). West and central Asia are large and humanly diverse regions which may contain uninvestigated parental groups of Europeans.

The EuroDNA 1.0 test results are correlated with phenotypic variations in hair, skin and eye pigmentation. These correlations may be too low to have useful predictive power for forensic uses or biomedical research but since the markers were not selected for that purpose it is promising that screening from large SNP panels may provide better power.

Is the observed dearth of European stratification with EuroDNA 1.0 markers conclusive? The next chapters will investigate this question through broader AIM screenings and more targeted sampling of Europeans and neighboring populations.

Chapter 3

Measuring European Population Stratification with Microarray Genotype Data

Introduction

This chapter is an expanded version of a manuscript published in the *American Journal of Human Genetics*[85]. This article reports on genome-wide typing of *ca.* 10,000 single nucleotide polymorphisms (SNPs) in 297 individuals to explore population structure in Europeans of known and unknown ancestry. The results reveal the presence of several significant axes of stratification, most prominently in a north-southeastern trend, but also along an east-west axis. The paper also demonstrates the selection and application of EuroAIMs (European Ancestry Informative Markers) for ancestry estimation and correction. The Coriell "Caucasian" and CEPH Utah sample panels, often used as proxies for European populations, are found to reflect different subsets of the continent's ancestry. This chapter reports additional AIMs for axes of ancestry in Europe and Eurasia, whenever enough individuals are present for AIM screening. Approaches for finding AIMs that are unique to specific axes of ancestry are also described.

Methods and Material

Population Samples

Details on sample collection by the Shriver Lab are given in the Methods section of chapter 2. Overall I selected 297 individuals from 21 European and world populations who were genotyped for *ca.* 10,000 autosomal SNPs, primarily using Affymetrix 10K (Santa Clara, CA) mapping array technology[86]. Except for the samples already published (Table 3.1), the genotyping was prepared by Shriver Lab students; hybridization and scanning was performed at the Penn State microarray core lab facilities. The European population samples selected represent a broad range of the geographic and linguistic diversity of the continent (Table 3.1). Briefly, they consisted of: western Irish (n=6), eastern English (n=8), French (n=1), German (n=8), Valencian Spanish (n=20), Basque Spanish (n=8), Italian (n=9), Polish (n=8), Greek (n=8), Finnish (n=7), Armenian (n=8) and Ashkenazi Jewish (n=5). The Italian, Ashkenazi Jewish, and Greek samples also include respectively 2, 1 and 1 individuals from the Coriell Cell Repository. For broader context, the European populations were examined together with two African population samples [Mende from Sierra Leone (n=22) and Burunge from Tanzania (n=20)] as well as several Asian populations [Brahmin (n=11) and Mala (n=11) from India, and Central Asian Altaian (n=20)]. One Middle Eastern and two North African individuals from the Coriell

panel were also included, all three also labeled as Arabs in inconspicuous entries of the Coriell database.

Table 3.1 – Individual samples description

Nation of origin / Ethnicity	Language Family / Subfamily	N	Origin / Collected by		Affymetrix GeneChip® Array
"Caucasian"	*Unknown*	42	Coriell Institute[42]		10K Xba 131
central France	Romance /Italic	1	M. Bauchet		
Valencia, Spain		20	E. Parra (University of Toronto)[34]		
Italy		2	Coriell Institute		
Italy (south and Sicily)		5	M. Bauchet, B. McEvoy., M. Shriver		10K 2.0 Xba
Utah (USA)	Germanic	74	CEPH[42]		100K
Hanover, Germany		8	R. Deka (University of Cincinnati)		10K 2.0 Xba
East England		8	B. McEvoy.		
Ashkenazi Jewish (US)		4	M. Bauchet, M. Shriver		
Poland	Slavic	8	M. Bauchet, B. McEvoy., M. Shriver		
Greece	Hellenic	7	B. McEvoy.		
		1	Coriell Institute		10K Xba 131
Basque region (France and Spain)	Basque	8	S. Alonso (University of the Basque Country, Bilbao, Spain)		10K 2.0 Xba
Connaught, Ireland	Celtic	6	M. Shriver, B. McEvoy		
Armenia (one person per province)	Armenian	8	T. Sarkisyan and K. Hovhannesyan		
Finland	Finno-Ugric	7	A. de la Chapelle (Ohio State University)		
Ashkenazi Jewish	Semitic	1	Coriell Institute		10K Xba 131
Middle East		2			
North Africa		1			
Mende (west Africa)	Niger-Congo	22	G. Argyropoulos (Pennington Biomedical Research Center)[34]		
Burunge (east Africa)	Cushitic	20	S. Tishkoff (U. Maryland)[34]		
Altai Republic (Central Asia)	Turkic / Altaic	20	T. Schurr (University of Pennsylvania, Philadelphia, PA)[34]		
Andhra Pradesh (India)	Indic	11	Brahmin (upper caste)	L. Jorde and M. Bamshad (University of Utah)[34]	
		11	Mala (lower caste)		

NOTE.—All samples described here were collected with appropriate human subject approvals from the various institutions involved and under the principle of informed consent. Samples already used in a previous paper were indicated by a reference in the Origin column, and were assayed directly by Affymetrix[34,42].

Some individuals were typed for 11,071 autosomal SNPs using the Affymetrix 10K Xba 131 array and others on the newer 10K 2.0 Xba array (Table 3.2). A total of 9,724 SNPs overlapped between the two platforms and these formed the core data for analysis. Two European-derived populations were subsequently included: the first of these, the Coriell "Caucasian" panel (n=42), curated by the Coriell Cell Repositories, was typed using the Affymetrix 10K Xba 131 array. This "Caucasian" sample has been used to portray European variation; for example, it was the core European representative sample in The SNP Consortium allele frequency project [77]. However, the

genetically and socially ill-defined term "Caucasian" leaves doubt as to which population(s) this sample represents and how well it does so. The second European proxy sample is the *Centre d'Etude du Polymorphisme Humain* (CEPH) panel composed of European-American Utah residents, sampled in 1980, with declared ancestry from northern and western Europe; this group forms one of the four populations used in the international HapMap project. The CEPH Utah panel used here is made up of 74 unrelated individuals from family trios, 32 of which overlap with the HapMap individuals labeled CEU (CEPH Europeans). Each CEPH individual was genotyped on the Affymetrix 100K Mapping array for *ca.* 100,000 SNPs but only the 6,207 markers overlapping with the 10K dataset were considered here. In an effort to minimize the effects of missing data, each analysis includes only SNPs that had genotypes for at least one individual in each population sample. This resulted in slightly different SNP sets for each comparison but average missing data rate per individual never exceeded 3.5%.

Table 3.2 – PCoA significance tests

	All 9,111 SNPs		SNPs >50kb apart (6,349 SNPs)		SNPs >100kb apart (5,555 SNPs)	
	Adj. R^2 (P)†	SKT P‡	**Adj. R^2** (P)†	SKT P‡	**Adj. R^2** (P)†	SKT P‡
PC1	**0.90** (<0.001)	<0.0001	**0.89** (<0.001)	<0.0001	**0.90** (<0.001)	<0.0001
PC2	**0.78** (<0.001)	<0.0001	**0.74** (<0.001)	<0.0001	**0.72** (<0.001)	<0.0001
PC3	**0.43** (<0.001)	<0.0001	**0.50** (<0.001)	<0.0001	**0.35** (<0.001)	<0.0001
PC4	**0.54** (<0.001)	<0.0001	**0.30** (<0.001)	<0.0001	**0.19** (<0.01)	<0.01
PC5	**<0.1** (NS)	NS	**<0.1** (NS)	NS	**0.13** (<0.05)	<0.001
PC6	**<0.1** (NS)	NS	**<0.1** (NS)	NS	**0.18** (<0.01)	<0.001
PC7	**<0.1** (NS)	NS	**0.17** (<0.01)	NS	**0.18** (<0.01)	<0.01
PC8	**<0.1** (NS)	NS	**<0.1** (NS)	NS	**<0.1** (NS)	NS
PC9	**<0.1** (NS)	<0.01	**<0.1** (NS)	NS	**<0.1** (NS)	NS
PC10	**0.13** (<0.05)	NS	**<0.1** (NS)	NS	**<0.1** (NS)	NS

NOTE.—Significance tests using the ANOVA and SKT. PCoA was conducted separately for the each SNP set, excluding the French singleton and the German outlier. Percentages of the variance explained by each PC are similar in all three SNP panels. NS means not significant at the 0.05 level.

† ANOVA correlation coefficient (adjusted R-squared) in bold with corresponding P-value in parenthesis. Coefficients <0.1 were not significant at the 0.05 level.

‡ Combined P-value calculated for SKT as described in the text

Detecting Relatedness

The presence of closely related persons can affect results and interpretation in unforeseen ways. Therefore I used a program performing maximum likelihood estimation of relationship between individuals (ML-Relate[71]) in order to identify closely related individuals. Because this program is

highly computationally intensive I used 484 unlinked SNPs from chromosome 1 (each SNP over 100kb apart). One Coriell Italian and one Coriell "Caucasian" turned out to be full siblings, and so did two Spanish individuals. For each sibling pair, the individual with the highest rate of missing data was excluded from subsequent analysis. Close but not first degree relationship was also suggested among several of the Coriell "Caucasians", hinting that they may be distant relatives.

Network Analysis

Neighbor-joining (NJ) trees of individuals[87] were used to analyze allele sharing distance (ASD) matrices (see Chapter 2, Methods) to graphically represent inter-individual genetic distances[34,42,88]. However NJ trees assume bifurcating topological relationships among populations, which may not represent well a continent where populations are inter-related in complex ways. This method was only used to confirm outliers and related individuals (not shown).

Principal Coordinate Analysis

Principal Coordinate Analysis[41] (PCoA) on inter-individual ASD matrices, methods to determine principal component (PC) axis significance are described in Chapter 2. In short they consist in the analysis of variance (ANOVA) test with each PC as dependent variable and population/group membership as predictor variable and a split-half reliability, the split karyotype test (SKT), which offers the advantage not to rely on population labels[34,42]. In order to ensure close marker spacing did not affect the results the SKT and ANOVA test were conducted on the PCoA results using the full SNP panel (9,111 SNP markers) and on two subsets including only SNPs at least 50Kb (6,349 SNPs) or 100Kb (5,555 SNPs) apart (Table 3.2 and Figure 3.1).

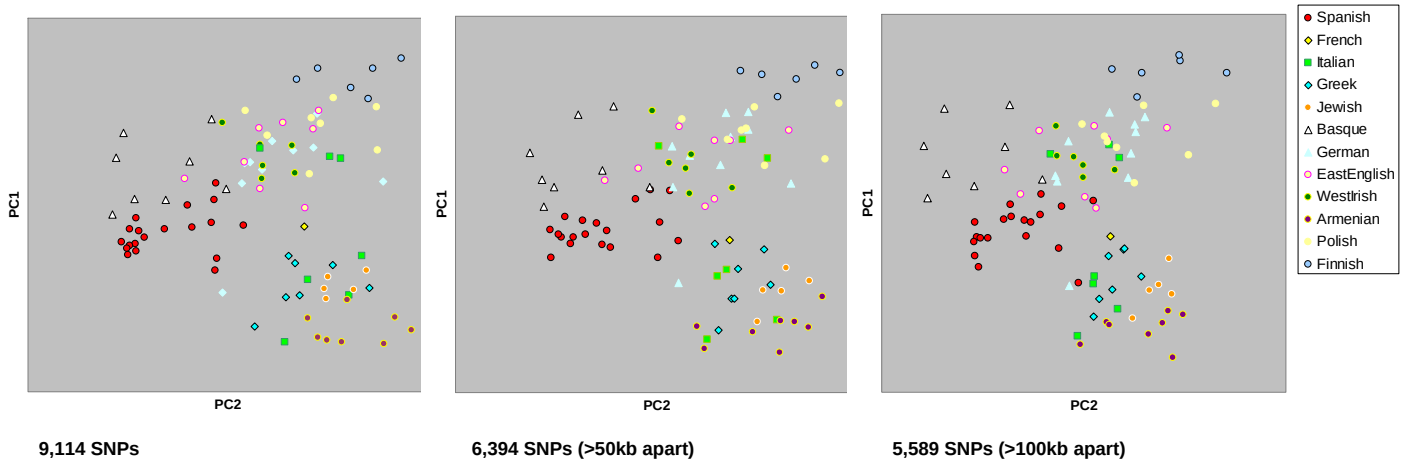


Figure 3.1 – Stability of the PC1-PC2 distribution of European individuals across marker sets with different minimum inter-marker separation (50Kb and 100kb).

Bayesian analysis

Bayesian analysis with the program *structure* 2.1[43,44] , allows fractional assignment of the genome to different populations. Each individual is represented as vertical line divided into at most K colored segments, where K is the pre-specified number of populations into which the data is to be divided. The program *structure* was performed directly on individual SNP genotypes and using the same parameters as in chapter 2. As previously, the *structure* plots were generated using the companion program *distruct*[83]. For the European population analysis the posterior probability (PP) is maximum at K=2. Indeed, the values of clusteredness (G)[4] are highest for K=3, then K=5 (Figure 3.2C) which correspond to two tentative stratification models described below.

Table 3.3 – European stratification models: Correlation between group membership and individual PC.

	ANOVA P-value		
	Model-0	Model-1	Model-2
PC1	<0.0001	<0.0001	<0.0001
PC2	NS	<0.0001	<0.0001
PC3	NS	NS	<0.0001
PC4	NS	NS	<0.0001
PC5	NS	NS	NS
PC6	NS	NS	NS

NOTE.—Models are based respectively on 2, 3 and 5 European clusters described in the Methods under the EuroAIMs Selection section. NS means not significant at the 0.01 level.

EuroAIMs Selection

The differentiation patterns observed in both PCoA and *structure* (Figures 3.2 and 3.5) lead to three models that describe the structure best; Model-0 with 2 clusters, Model-1 with 3 clusters and Model-2 with 5 clusters (Figure 3.2B). For each model, in order to minimize the effect of admixed individuals, individuals were chosen to represent a cluster when they had a stable high level in that cluster across multiple runs of *structure*. Then, using these representative individuals, I calculated F_{ST} for all SNPs between each cluster pair, and selected SNPs with the highest F_{ST} as potential EuroAIMs.

Consistent with simulations described in the previous chapter (Methods) the number of top PCs showing insignificant ANOVA tests is directly proportional to the number of clusters assumed (Table 3.3), further confirming the relevance of each model described below.

Model-0 – This model is purely based on PC1 and the *structure* run with the highest PP (K=2). This represents the dominant trend which separates the northern individuals from the southern. I identified EuroAIMs from the original panel of Europeans (Figure 3.2) by defining northern (n=36) and southeastern (n=31) cohorts of individuals based on extreme polar values in PC1 (above 0.5 or below -0.5). The northern cohort included all Finnish, Polish, most German, Irish and English, as well as some

Basque and Italian individuals. The southeastern cohort included all Armenians, Jews, Greeks and the other Italians. Weir's F_{ST} was then calculated for each SNP as a measure of genetic distance between the two groups. All SNPs were ranked by F_{ST}, with those showing the highest values likely to represent the best north-southeastern EuroAIMs (Figure 3.4). Although useful to determine north-south AIMs this model ignores the important east-west axis (Figures 3.2A and 3.3, PC2).

Model-1 - Three clusters are based on Figures 3.2B (PC1 and PC2) and 3.2C (K=3 has the highest clusteredness):

Cluster 1A: Basque and Spanish individuals

Cluster 1B: Greek, Armenian, Jewish, Sicilian and Coriell Italian individuals

Cluster 1C: German (except outlier), Polish, Irish, English and two south Italian persons

Model-2 – Five clusters can be distinguished on Figure 3.2C, where K=5 represents the second peak of clusteredness). This model is further justified by the high significance of PC3 and PC4 in the European dataset which respectively emphasize the separation of Basque and Finnish individuals (Table 3.2 and Figure 3.3):

Cluster 2A: The Finnish cluster contains all Finnish individuals

Cluster 2B: Germans (except one), Poles, Irish, English, 2 mainland Italians

Cluster 2C: Spaniards from Valencia

Cluster 2D: Basques (distinguished by PC3)

Cluster 2E: Greeks, Armenians, Jews, Sicilians and Coriell Italians

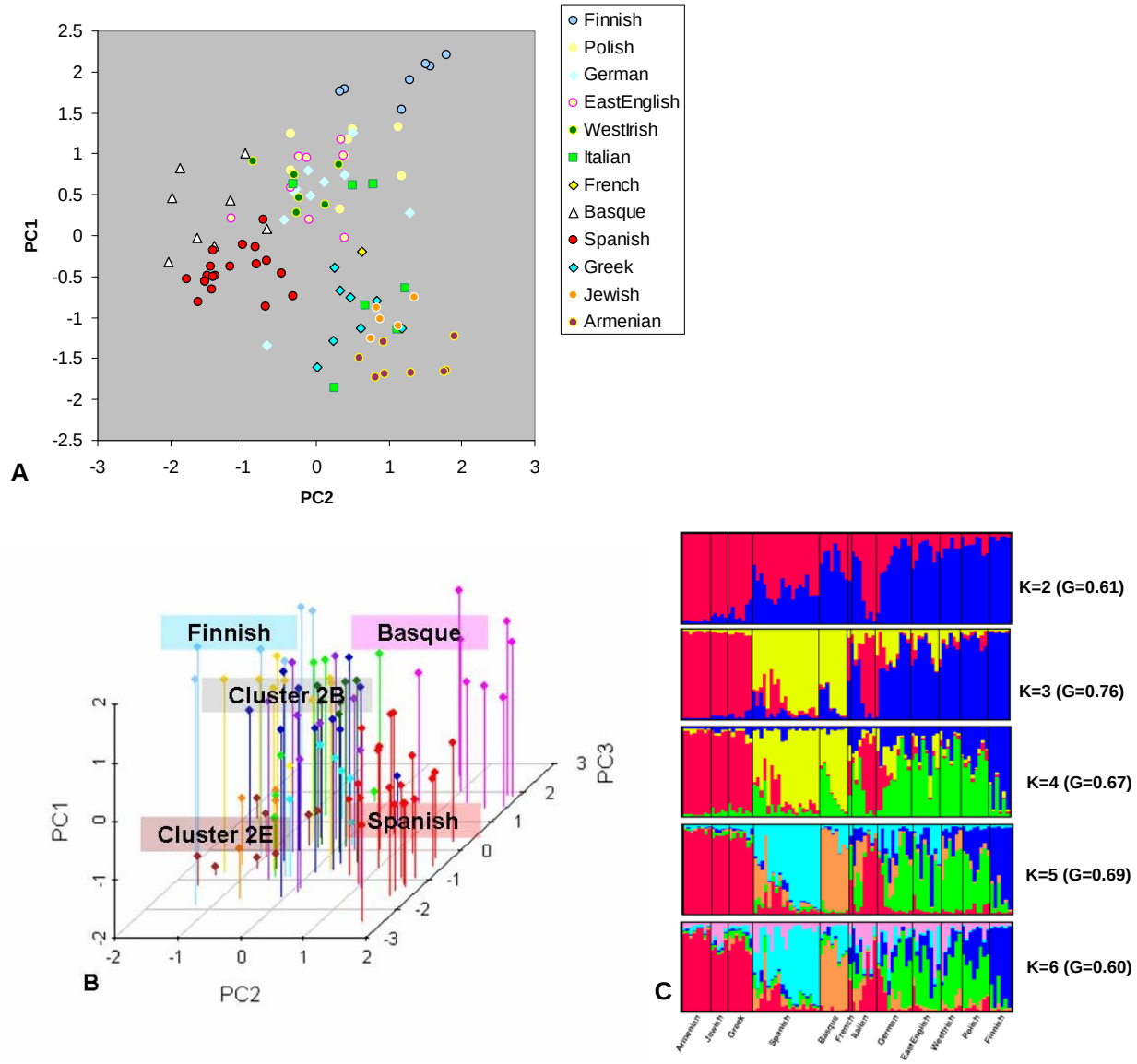


Figure 3.2 – Population structure in European individuals. ***A,*** Geographically oriented PC 1 and 2 putting forward the 3-cluster model (Model-1) and ***B,*** adding PC 3 for a 3D representation of 5 putative clusters for Model-2. PCoA is based on average inter-individual ASD over 9,114 SNPs. PC 1, 2, 3 and 4 (Figure 3.3) respectively explain 2.05%, 1.7%, 1.6% and 1.5% of the variation and were highly significant by SKT and ANOVA testing (Table 3.1). ***C,*** Bayesian clustering analysis using *structure* with the same markers and European individuals.

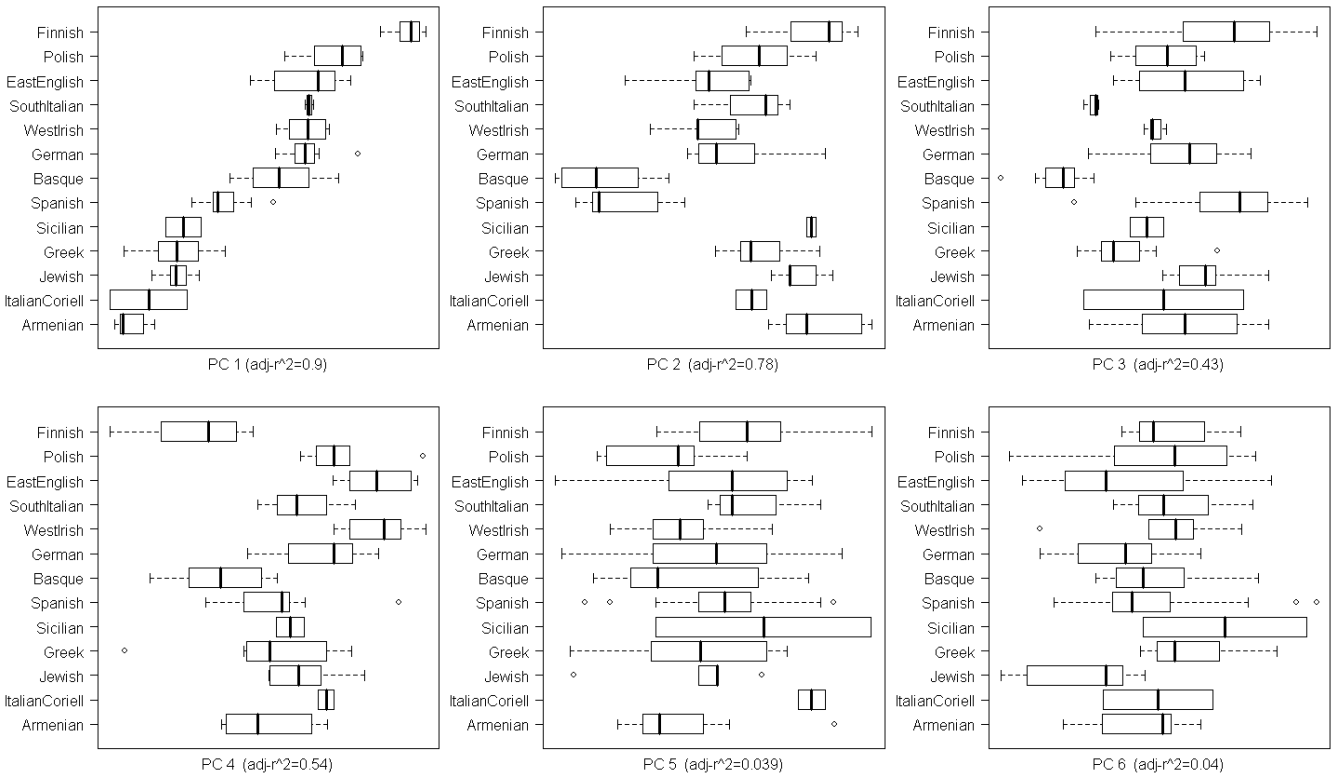


Figure 3.3 – PCoA plots for the first six PCs in European samples. The bolded vertical bar represents the median PC value of each group, the two hinges are the first and third quartile and notches give an approximate 95% confidence interval for the difference in two medians. Adjusted R^2 is from the ANOVA test.

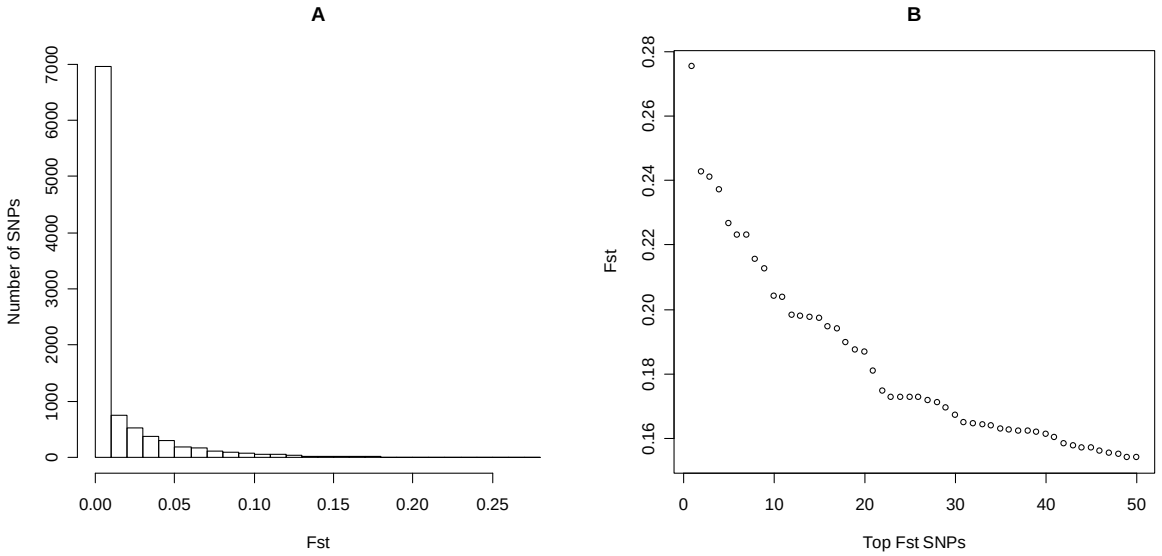


Figure 3.4 – Distribution of F_{ST} between northern (n=36) and southeastern (n=31) cohorts of individuals selected from PC1 values in Figure 3.2 (above 0.5 or below -0.5). ***A,*** Histogram using all 9,721 SNPs available. ***B,*** Plot of top 50 SNPs of highest F_{ST} (see also Table 3.4).

Measuring Stratification

In order to assess the potential impact of the observed European stratification in case-control association studies, I calculated the factor by which association statistics might be inflated (how much more likely false positives are to arise due to population stratification)[89]. A simple estimator for this inflation factor λ is the mean allele frequency correlation between cases and controls (χ^2) across null-loci (loci not thought to influence the trait or condition)[39]. I examined the most extreme scenario supported by our data, where the case and control groups are composed of northern and southeastern individuals respectively. I simulated 1,000 cases and 1,000 controls based on these cohorts' observed allele frequencies and calculated the mean χ^2 across all loci at least 50kb apart with an allele count of at least 5 (6,312 SNPs). Finally the λ estimate is obtained by multiplying the result by 1.03 which in this case is the maximum factor by which λ can exceed the mean χ^2 at the 95% confidence level[39].

Statistical Analysis

Computations were performed and most figures were generated using scripts as indicated in Chapter 2.

Results and Discussion

Europe and Neighboring Continents

The PCoA clearly identifies four widely dispersed groupings corresponding to Europe, South Asia, Central Asia and Africa (Figures 3.5A, 3.5B and 3.6). In these figures, PC1 appears to separate the two African populations from the others while PC2 divides the Asians from the Europeans and Africans, and PC3 splits the Central Asians from the South Asians. A complementary Bayesian approach using the program *structure*[43,44] supports the PCoA findings (Figure 3.5C). When the number of putative populations is set at four (K=4), the groups largely correspond to the same four regional divides apparent from the PCoA.

An overview of Europeans in a continental context includes all Eurasian individuals and Africans (Figures 3.5 and 3.6). Four main continental clusters are observed in Figure 3.2:

1- "African" meta-cluster, clearly structured into two sub-clusters (Mende and Burunge)

2- "South Asian" meta-cluster, including two sub-clusters (Brahmin and Mala)

3- "Central Asian" cluster (Altaians)

4- "European" cluster (which the 3 Arab individuals are closest to)

Overall these form six non-overlapping clusters and sub-clusters, consistent with *structure*'s highest posterior probability (K=6).

The wide gaps observed between the four main clusters likely reflects the absence of geographically intermediate groups in the analysis, an explanation hinted at by the intermediate position of the North African and Middle Eastern subjects include (all three are also labeled as Arabs in alternate Coriell database entries; Appendix B). These three persons are too few to form their own *structure* cluster but their PCoA position suggests these region need to be explored further. It is interesting to note that both in *structure* (Figure 3.5C) and PCoA (Figures 3.5A and 3.5B) these persons show the strongest affinity with the European cluster, and the next strongest toward the African samples rather than either of the Asian ones. It is unlikely that these Arab individuals are from populations recently admixed from European and sub-Saharan African populations; rather, they are more likely to be closer genetically to the original progenitors of some or all Europeans, while maintaining a more African genetic background than Europeans because of gene flow due to geographic proximity. More samples and a better geographic representation from North Africa and the Middle East are needed, but it is unclear whether it would produce a smooth continuum or series of clusters between the groups observed here.

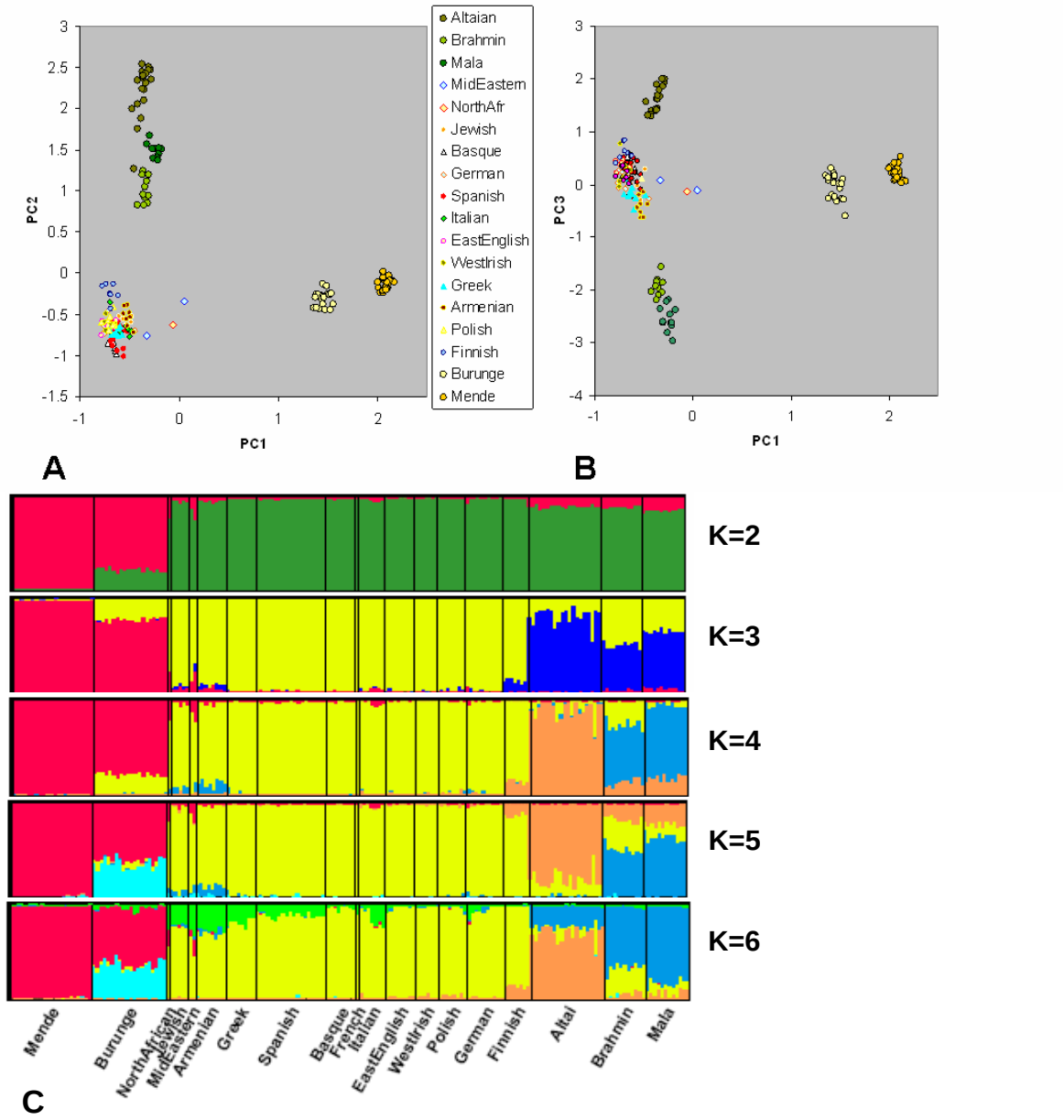


Figure 3.5 — Population structure in European, African and Asian individuals. ***A and B,*** PCoA results based on average inter-individual ASD using 9,100 SNPs. PC 1, 2 and 3 respectively explain 11.6%, 3.4% and 1.4% of the variation. ***C,*** Bayesian clustering results using *structure*[43,44] with the same markers and individuals.

It is also clear that geography is not the sole marker of differentiation. The South Asians Mala and Brahmin are low and high caste groups respectively from the same region of India (Andhra Pradesh) yet still show significant separation supporting some degree of socially maintained stratification and possibly different historical origins (Figures 3.5 and 3.6). Even though they display some common ancestry and admixture in *structure* (Figure 3.5C), the two castes do separate in PCoA more sharply than any two geographically separated European populations (Figure 3.5A and 3.5B).

The first principal coordinate (PC1) of Figure 3.5A and 3.5B is very consistent with the *structure* components present of Europeans and Africans (Figure 3.5C, K=3 and 4). Interestingly PCoA places the east Africans (Burunge) directly between west Africans and Europeans. One interpretation is that those specific east Africans are direct descendents of the primary progenitor population of the first human expansions out of Africa. However this is not suggested by this evidence because the Burunge are only located on one axis, toward the European cluster, and not toward any of the Asian clusters (Figure 3.5A and 3.5B). Other possibilities are persistent gene flow between Europe / Middle East and the eastern part of Africa, or that the Burunge descend directly from a population ancestral to the other groups. A denser population sampling in and around East Africa as well as the huge and ethnically diverse area of the Middle East would help to resolve these issues.

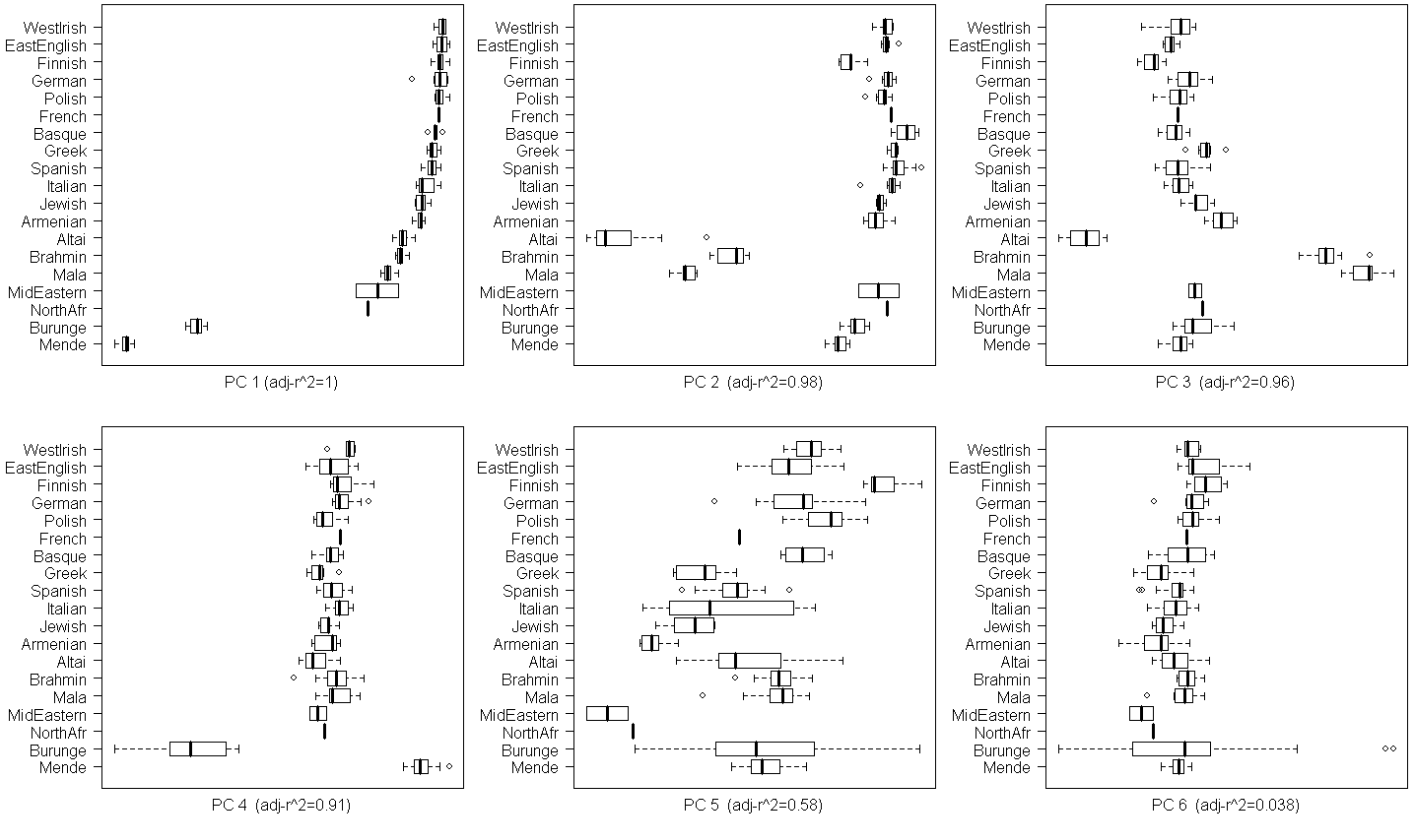


Figure 3.6 – PCoA box plots for the first six PCs in samples from Europe and neighboring continents. The bolded vertical bar represents the median PC value of each group, the two hinges are the first and third quartile and notches give an approximate 95% confidence interval for the difference in two medians. The overall correlation between group membership and PC value is reported by ANOVA's adjusted R^2 for each PC. The few subsequent PCs which are also significant pertain to Europe and are best observed in Figure 3.2 and 3.5.

Europeans showing the most affinity to the African clusters are the Greeks, Jews and Armenians (cluster 1B, or 2E), which however are revealed on PC3 to veer off towards South Asians (whereas the three Arabs keep their orientation towards Africans). On the other hand, the Finns, and to

a lesser extent the Poles, show clear affinities with Central Asians. Finally, the Iberian groups (Spanish and Basque) do not show more affinity toward the African meta-cluster than other Europeans, and of all European groups, have the least affinity with Asians. These observations are consistent with idea suggested by other sources that the Iberian Peninsula holds the most direct descendants of the pre-Neolithic migrations into Europe[11,90]; and with the hypothesis of a Franco-Cantabrian refugium holding populations in relative isolation until the end of the last glacial maximum ~15 kya[91].

In line with previous studies,[3] the observed diversity in Europe is low, with the entire continent-wide sample only marginally more dispersed than single population samples from elsewhere in the world.

Inside Europe

I next investigated European individuals in more detail using a similar approach. An initial Mantel test[92] between matrices of inter-individual geographic and genetic distances was highly significant ($P<0.001$) suggesting some degree of geographic sub-structure despite the relatively limited diversity. Since the clustering of individuals in PCoA space is dependent on which persons are included I restricted the analysis to individuals clearly attached to the European cluster using additional measures of significance to describe the more subtle patterns of structure within-Europe. The European meta-cluster happens to be composed of all geographically defined European individuals, as well as Armenians and US Ashkenazi Jews, both in PCoA and *structure* (Figure 3.5). Within these Europeans, one axis is pointing at Central Asia the other at the Middle East; the intersection of those two axes falls approximately in the Iberian groups (Figure 3.2A), approximately reflecting the geographical distribution of these populations. The first four PCs were found to be consistently significant in both the SKT and ANOVA test (Table 3.2), and thus are likely to represent real structure. The European samples either overlap or neighbor each other closely, but no group differences are as wide as observed at the Eurasian level on any of the four significant PCs. PC1 largely separates northern from southeastern individuals (Figure 3.2A) and is consistent with the clines observed in classical genes frequencies,[11,52] Y chromosome,[31] mtDNA[30,93] and recent whole-genome[46] studies of European diversity. PC2 reflects mainly east-west geographic separation and particularly identifies Iberian individuals as distinct (Figure 3.2A). PC3 distinguishes the Basques from all other Europeans and PC4 emphasizes the separation of the Finns already noticeable along PC1 (Figure 3.2 and 3.3). The Basques are known to have unusual allele frequencies for several marker systems[94] and speak a unique non-Indo-European language. The Finns speak a non-Indo-European Uralic language and in line with this

fact and previous Y-chromosome work,[95] show evidence of an increased affinity to the Central Asian populations when placed in an inter-continental context (Figure 3.5A and 3.5B).

The stability of the PCoA findings across different marker separation sets (50Kb and 100Kb, Figure 3.1) suggests that geographic structure is distributed throughout the data, and that nearby markers in the 10K arrays are redundant in terms of ancestry informativeness. The slight degradation and dilution of the significance towards higher PCs (in particular in the >100kb case, Table 3.2), is consistent with lower resolution power due to fewer markers. The main point here is that linkage does not create spurious patterns but if anything will likely improve the sharpness of clustering, both in PCoA and *structure*.

Overall, *structure* analysis is highly consistent with PCoA; Model-1 is the run of highest PP, and Model-1 and Model-2 are the two runs of highest clusteredness (Methods, EuroAIMs Selection).When the number of clusters is three (K=3), the major divisions correspond to the Model-1 of northern, southeastern and Iberian populations (Figure 3.2B). In cases of higher Ks, first the Finns (K=4), then the Basques (K=5) emerge as distinctive, as reflected by Model-2.

Table 3.4 – Top 20 northern-southeastern EuroAIMs (Model-0)

dbSNP ID	Chromosome	Weir F_{ST}	Delta[a]
rs988436	5	0.2755	0.295
rs942793	10	0.2428	0.342
rs1368136	8	0.2412	0.379
rs2060983	8	0.2373	0.377
rs4988235 [b]	*2*	*0.2352*	*0.374*
rs1404402	1	0.2267	0.354
rs1016120	2	0.2232	0.269
rs1414411	1	0.2232	0.365
rs2014303	4	0.2156	0.332
rs1030626	8	0.2126	0.355
rs1517661	12	0.2041	0.348
rs764138	16	0.2039	0.349
rs2218497	13	0.1981	0.345
rs725379	2	0.1980	0.338
rs1377724	15	0.1974	0.230
rs1406121	2	0.1973	0.345
rs869538	4	0.1945	0.309
rs1905471	13	0.1940	0.236
rs764681	16	0.1898	0.321
rs1280100	4	0.1873	0.320
rs723211	10	0.1867	0.333

NOTE.—The full set of 1,200 EuroAIMs between the northern and southeastern cohort is available at Web Resource 2.
[a] Allele frequency difference between cohorts.
[b] This SNP is not in our original set but was part of a study where it showed significant stratification between northwestern and southeastern Europeans.[37]

Correction for stratification in association studies is dependent on the identification and adjustment for relevant axes of ancestry namely those that vary within the study population. While a large number of arbitrary SNPs can be used for this purpose, a more efficient and informative approach is to identify subsets of ancestry informative SNPs, such as EuroAIMs. I focused here on the main recognizable trend in this European data set, i.e. PC1 or the north-southeastern ancestry axis (Model-0). The 20 SNPs presenting the highest F_{ST} levels between the northern and southeastern cohort are listed in Table 3.4 along with allele frequency differences between cohorts (Methods, *EuroAIMs Selection*). This table also shows that these top 20 EuroAIMs have levels of divergence comparable to a SNP (*rs4988235*) which stratification was previously shown to induce false positives in European populations along the same geographical axis[37].

Measuring stratification for PC1 as described in Methods, Measuring Stratification yields a maximum inflation factor $\lambda_{max} \approx 48$, substantially exceeding the null hypothesis expectation of 1 (zero stratification). In other words, a conservative P-value for a candidate SNP can be obtained after dividing the χ^2 value (or any association statistic) by 48.[39] This confirms the practical importance of PC1 stratification in European populations and the utility of the selected EuroAIMs panels to control for it. Therefore these marker sets were further tested in PCoA and *structure* analysis of two European-derived population samples; the Coriell and CEPH individuals.

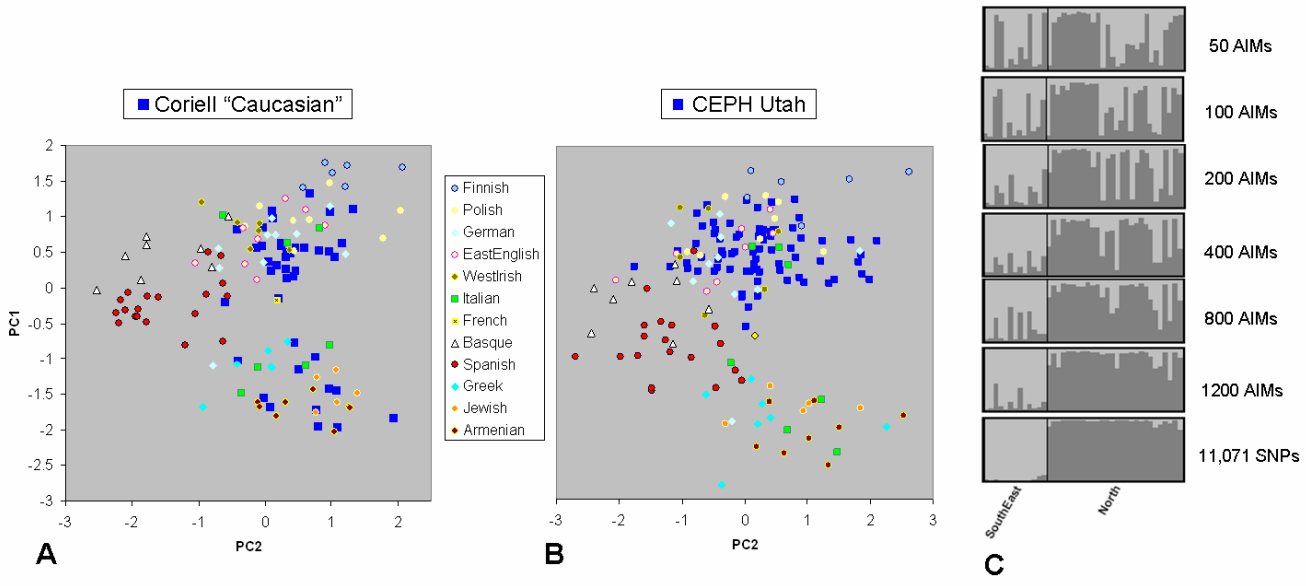


Figure 3.7 – Population structure in panels of European-derived ancestry within the context of European individuals (from Figure 3.2A). ***A,*** PCoA of the Coriell "Caucasian" panel (n=42) together with Europeans of known ancestry, based on all 9,114 SNPs in common. ***B,*** PCoA of the CEPH Utah individuals (n=74) and Europeans using all 6,207 SNPs in common. ***C,*** *structure* runs using the Coriell "Caucasian" sample based on the full SNP dataset (bottom) as well as sets of different number of north-southeast EuroAIMs (Appendix C).

Stratification in European-derived samples

All European genotypes mentioned above were first put together with the Coriell "Caucasians" and the same analyses were repeated. One Coriell "Caucasians" (NA17205) already detected as outlier in a previous genomewide study of worldwide individuals[42], appeared closest to (but not within) the Brahmin cluster. As this person may likely be of mixed western European and Indian ancestry of from a geographically intermediate region of west Asia, he was excluded from this European specific analysis.

PCoA analysis of the Europeans together with the Coriell "Caucasian" panel (Figure 3.7A) reveals the latter to be divided by the same north-southeastern structure evident in the general sample. The Coriell "Caucasians" falling in the southeastern cluster are widely distributed in this cluster; one of them turned out to be related to a Coriell Italian, which is consistent with the presence of Sicilian Italians in this cluster. Although the Coriell "Caucasian" sample clearly derives from at least two European populations, it lacks the full range of variation observed in Europeans, specifically that observed in the Spanish, Basque, and Finnish individuals. north-southeastern stratification is also evident in *structure* analysis based on the full set of SNPs (bottom of Figure 3.7C) and was used to classify the Coriell "Caucasian" panel into a north group and southeast group, which correspond to the two observed PCoA clusters (Figure 3.7A). This structure may largely be captured using 10 fold fewer SNPs provided these markers are EuroAIMs selected as most informative on the north-southeastern axis from the full 10K set. Using fewer than 1,200 EuroAIMs of the type available in this panel gradually leads to loss of consistent structure and a corresponding increase in misclassification of subjects (Figure 3.7C).

As an example of how AIMs could be used to determine ancestry in a sample of unknown European-derived ancestry, I used the Coriell "Caucasian" sample. The question becomes then; is it possible to determine along which axes of European stratification they fall, using a reduced set of markers? I attempted to select sets of EuroAIMs informative for each pair of cluster, alternatively in Model-1 (Table 3.5) and Model-2 (Table 3.6)

Table 3.5 - Model-1, SKT Combined P-values for PC1 (next 5PCs never significant). F_{ST} cutoff of 0.10 yields between 269 and 380 AIMs per cluster pair.

	Cluster 1B	Cluster 1C
Cluster 1A	0.99	0.105
Cluster 1B		**7.8e-29**

From SKT in Model-1 (Table 3.5) we can conclude that the Coriell "Caucasian" panel is stratified between Cluster 1B (southeastern) and Cluster 1C (northern), but not for any other cluster pair—no significant structure was found to be caused by individuals from Cluster 1A (Iberians). The PCoA plot confirms this (Figure 3.7A), but also suggests that further refinement of the structure present in this "Caucasian" sample using Model-2.

Table 3.6 - Model-2, SKT Combined P-values for PC1 (next 5PCs never significant). F_{ST} cutoff of 0.15 yields between 158 and 833 AIMs per cluster pair.

	Cluster 2B	Cluster 2C	Cluster 2D	Cluster 2E (SE)
Cluster 2A (FN)	1	0.002	2.8e-06	1.1e-016
Cluster 2B (NW)		0.986	0.972	**< e-200**
Cluster 2C (SP)			0.998	4.6e-008
Cluster 2D (BQ)				1.5e-14

In Model-2 (Table 3.6), in agreement with PCoA using all markers (Figure 3.7A), the Coriell "Caucasian" appears mostly stratified between Cluster 2B (northwestern) and Cluster 2E (southeastern) because that is where the combined P-value appears to be the most significant (<e-200). Other significant signals appear as well, and might best be considered false positive due the large overlap between AIMs sets, i.e. many SNPs happen to be AIMs for several cluster pairs (633 AIMs had $F_{ST} > 0.15$ in more than one cluster pair). My attempts to find "private" AIMs (i.e. AIMs specific for particular cluster pairs) eliminated too many high F_{ST} markers and yielded AIMs sets that were too small and with SNPs of too low F_{ST}. Therefore the difficulty in establishing a proper cutoff to the combined P-value in Table 3.5 may be due to the lack of power of the available 10k markers in resolving differences among clusters of Model-2.

Finally, the CEPH Utah individuals cluster with northwestern Europeans in line with their more restricted described origins within Europe (north and west) and represent only a fraction of the north-southeastern variation (PC1) observed in the Coriell "Caucasian" panel (Figure 3.7B). Furthermore, despite the dispersion of CEPH individuals along PC2, the SKT did not detect any significant stratification along this axis. Bayesian analysis with *structure* also failed to detect meaningful structure either using EuroAIMs or the full set of available markers for this dataset (6,207 SNPs). The substantially higher levels of stratification seen in the Coriell sample relative to the CEPH European American sample supports previous conclusions regarding those samples[34].

The top 1,200 north-southeast (PC1) EuroAIMs can be downloaded from the Shriver Lab website (http://www.anthro.psu.edu/biolab/euroaims.pc1.xls). Using F_{ST} in the same way the top 1,200

east-west EuroAIMs were selected along PC2, but lack of additional independent human samples prevented testing these AIMs (http://www.anthro.psu.edu/biolab/euroaims.pc2.xls).

Conclusion

Both analytical approaches, Bayesian analysis with PCoA on ASD matrices and *structure*, present striking concordance in clustering individuals and their graphical representations complement each other. The varied individual *structure* profiles cannot always easily be interpreted as admixture, primarily because this survey doesn't cover all European regions and European possible demographic sources. Some regions, such as North Africa and the Middle East are represented by too few individuals to form a *structure* cluster, although their admixture levels and position in PCoA space suggest strong European affinities (Figure 3.5). Bayesian analysis with *structure* has the advantage of giving admixture estimates, but interpretation requires caution because any individual seemingly admixed may either be truly admixed, or share common ancestral roots with individuals geographically and historically intermediate between two clusters. Individuals may appear admixed in *structure* for several reasons, including because there are fewer individuals than in the other samples, or because they share common origins. For instance in Figure 3.5C the Mende have almost 100% clusteredness whereas the Burunge appear admixed with European. However it is equally likely that the Burunge represent a parental population to both Mende and Europeans. PCoA plots help alleviate some of the ambiguity, and if not, they point to the possible absence of key related populations (parental or intermediate) such as west Asia, eastern Europe and North Africa in Figure 3.5. The importance of contextual populations in understanding European variation is illustrated by the example of the Finns, who appear very distinct and possibly parental to some European populations in some *structure* runs (Figure 3.2A). However, from a Eurasian perspective it becomes apparent that Finns are also likely to be intermediate or admixed with populations further east (Figure 3.5). A similar phenomenon can be observed with Armenians which are located somewhere along a genetic gradation between South Asians and western Europeans.

Two potentially useful ways to describe these multidimensional clustering patterns were devised in order to extract the most useful EuroAIMs. Model-0 (north-southeastern) and Model-1 (3 cluster, Figure 3.2A) appears to be the most useful in yielding EuroAIMs, whereas the more refined Model-2 (5 cluster, Figure 3.2B) was unable to provide enough useful EuroAIMs.

The Iberian Peninsula has at least 3 main historical population sources; 1) prehistorical Iberian populations, whose diversity and structure is unknown, 2) the Celts (originally mostly from Central Europe), merging with some of the Iberians to form the Celtiberian civilization, 3) the Roman invaders[1]. Here the Basque and Valencian Spanish each form a cluster on their own but tend to form together a meta-cluster distinct from other Europeans, suggesting the continuity and diversity of a distinct Paleolithic (pre-Roman, pre-Celtic) Iberian ancestry. The Celtic influence may explain why some Spanish subjects appear to drift toward the northwestern Europeans, although, considering Italian diversity (discussed below), the Roman invasions may well have also brought along a significant number of "Cluster 2E" people to Iberia. Additionally, any or all Spanish individuals may also have Basque admixture as suggested by *structure* run Figure 3.2C (K=5). Additional Iberian and French population samples may well allow a more precise picture.

In some cases, geographic distance or physical barriers are not well reflected in these results. For instance, despite their insular origin, Irish and English individuals cluster with the continental Germans and Poles. Similarly, large geographical distances such as between Greece and Armenia are smaller at the genetic level. Conversely, Italy appears to be a zone of sharp differentiation over small distances. However the north vs. south gradient previously observed with classical markers[11] may not be the most important trend at the genome level. Even though more than these 7 Italian individuals are needed for a definite conclusion, it is of note that the 3 Italians clustering with the northern Europeans are from south Italy, while the others falling into the southeastern grouping are Sicilians and Coriell Italians (Figure 3.2A). The SKT confirms significant stratification within both Cluster-2E (northern) and Cluster-2B (southeastern) and within the Spanish sample. However, Mantel correlations between genetic and geographic distance were not significant within Cluster-2E and Cluster-2B[85]. It is likely that additional populations, additional individuals for some populations and an increased number of markers will be required to investigate the nature and extent of these more subtle patterns—as well as precise information on ethnic origins.

The 8 Armenian individuals are the furthest on the South Asian axis of ancestry. They form a meta-cluster with the Greek, Sicilian and Ashkenazi Jewish individuals (Cluster 2E), which is remarkable considering the wide geographical gap between Armenia in the Trans-Caucasus and the Mediterranean populations with whom they are so close genetically. This is highly consistent with the linguistic theory that the Armenian, Hellenic and Iranian languages together form a subgroup of the Indo-European language family[96]. However the same genetic cluster includes Jewish individuals whose ancestors spoke a non-Indo-European Semitic language. The one Jewish individual from the Coriell Ashkenazi panel clusters consistently with the four others (US Ashkenazi Jews we collected,

with all four Ashkenazi grandparents) in the midst of Cluster 2E, which suggests little admixture occurred from the northern European populations among which they were living for many centuries—although more individuals are needed from this population since they may have been undergoing different demographic events in various parts of Europe.

Principal components analysis using classical markers have identified different directional gradients[22,28], mainly following a southeast to northwest cline interpreted as the Neolithic population expansion following the rise of agriculture and sedentism[11,23,24]. However, the original peopling of Europe in the Paleolithic, which is one migration known to have followed a model of demic diffusion[16,17,97], likely followed the same routes as the Neolithic expansions. Lack of knowledge on the genetic composition of ancient migrations prevents us from drawing definite conclusions from the observation of the current genomic landscape. However, from a practical perspective this genome-wide investigation of European ancestry, in combination with other recent studies[36,37,98], demonstrates the importance of considering population stratification in studies using European and European-American individuals. Further examination of additional population samples, more individuals per population and a larger number of markers will allow refinement of the important axes of variation. This will in turn enable the selection of efficient EuroAIMs sets for measuring and correcting for stratification within European-derived population samples, and will also inform the debate on the population history of Europe.

To which level of detail can we describe stratification if we include all the existing human genetic variation? The HapMap project is limited in having only one European population, which only represents a partial amount of European genetic variation (chapter 2). It is currently unrealistic to genotype individuals for as many markers as HapMap but the following Chapter 4 presents a manifold increase of SNP numbers in comparison with the present analysis.

Chapter 4

European Populations and Individuals in Light of 317k SNPs Genotyping

Introduction

The use of Illumina 317k mapping arrays (317,503 SNPs), first on 29 population samples from Europe and neighboring continents, then on 180 European individuals (mainly from Ireland, Italy, Poland and Portugal) allows us to gain further qualitative and quantitative insights into the axes of genetic stratification in Europe.

First, an efficient way to make use of high-throughput mapping arrays is to pool individuals together on a single array[99]. With the knowledge gained on European population genetic landscape from smaller mapping array genotyping (10k, Chapter3) as well as population history and culture, I collaborated with researchers at Stanford University (Palo Alto, CA) to design a set of population pools (Table 4.1, Figure 4.1) that they genotyped on the Illumina 317k mapping arrays. First, I present and discuss a multivariate analysis of this dataset, as well as its use in screening ancestry informative markers for Europe (EuroAIMs).

In a second step, 180 persons (Table 4.2) were selected to be genotyped individually on the same 317k mapping arrays. This sample consists of 45 female individuals sampled in each of four cities of Europe, namely Dublin (Ireland), Rome (Italy), Warsaw (Poland) and Porto (Portugal). The square distribution of the geographic locations of these cities offers a minimal spatial sampling of Europe (Figure 4.1). The multivariate analysis of this dataset is also presented here, with the further screening for EuroAIMs using those selected in the first step with the pools. These 180 individuals were measured for several phenotypes, including skin, hair and eye pigmentation.

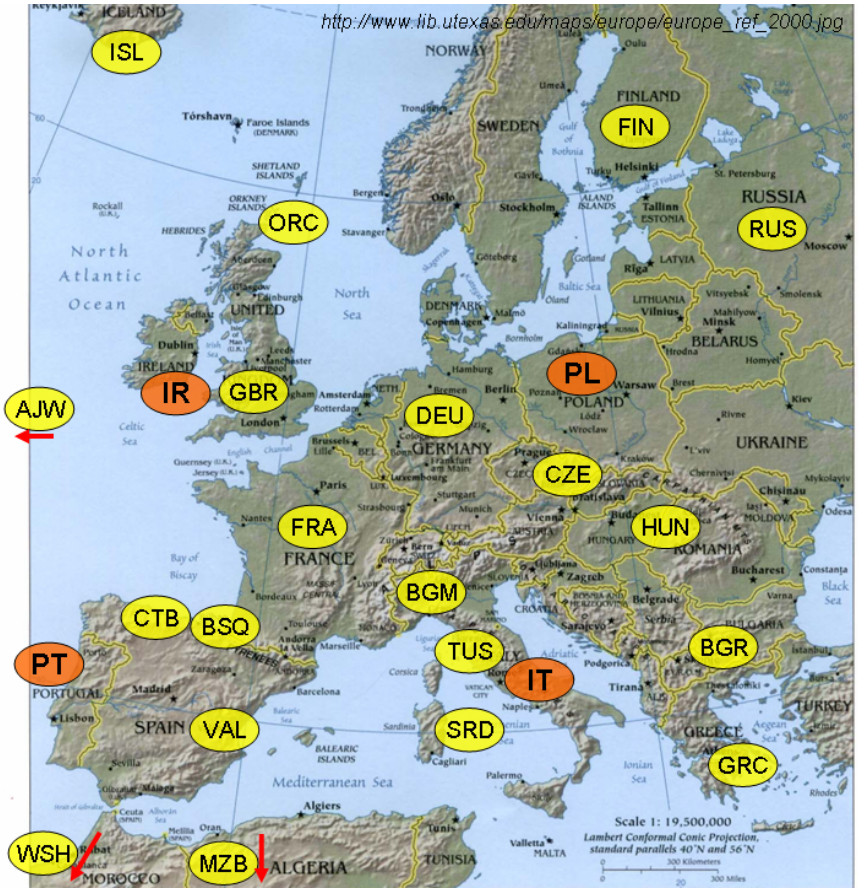


Figure 4.1 – Population samples genotyped on Illumina 317 mapping arrays. Pool samples are indicated in yellow circles with the sample abbreviations from Table 4.1. Sampling sites of genotyped individuals are represented in orange circle; IR=Dublin (Ireland), IT=Rome (Italy), PL=Warsaw (Poland), PT=Porto (Portugal), ***A,*** Europe and North Africa, ***B,*** Eastern Europe and Central Asia.

Methods and Material

317k genotyping

The 317k Illumina mapping arrays allow genomewide genotyping for 317,503 SNPs throughout all chromosomes, using the Infinium technology[100]. Samples from the Shriver lab were prepared and sent by E. Quillen and L. Pearson. All genotyping was performed by our Stanford collaborators (R. Myers, G. Barsh, J. Li, D. Absher). Two types of genotyping can be performed with the 317k Illumina mapping arrays; either individuals of known ethnicity can be pooled together on one mapping array, or a single individual can be typed on one array. When pooled genomic DNA is used the concentrations have to be measured very accurately so that each individual is represented equally; then the fluorescent intensities are proportional to allele frequencies. In the case of one individual typed on one array, intensities indicate the genotype of each SNP.

Population samples

The 29 population pools genotyped on the 317k mapping arrays include three populations from Central Asia and two from northwestern Africa, the rest being from a wide range of European countries (Table 4.1). Pool results from populations in the same region (two German population samples, two Basque and two Russian). Once it was verified that they each were strongly consistent in PCA (not shown) their allele frequencies were averaged together. In the other countries with multiple population samples, namely Italy and Spain, the samples showed substantial dissimilarity among each other and were therefore kept as separate pools. Even though the ideal sample size for a mapping array pool is ~50[99], some populations fall short of this number due to availability. Bias introduced by small sample size is visible in multivariate analysis (Figure 4.2C) and will be discussed below.

The English sample was mainly taken in South Wales and includes both English and Welsh individuals—to a lesser extent (G. Kirov, personal communication).

The Russian 2 sample is mainly from the Novosibirsk area in Western Siberia (T. Schurr, personal communication).

The Armenian sample consists in six individuals from each of the eight regions of Armenia (K. Hovhannesyan, personal communication); this is the complete set from which the eight Armenians typed on 10k mapping arrays (Chapter 3) were drawn.

Sahrawis are mixed Arab-Berber people (speaking both languages), descending from Yemeni Arab tribes who migrated into Western Sahara about one thousand years ago (F. Leyva-Cobian, personal communication).

Table 4.1 – Population samples pooled on the Illumina 317k mapping array

Ethnicity	Geographic origin	Abbreviation	N	Provenance
Adygei	Russian Caucasus	ADG	17	HGDP-CEPH
Armenian	Armenia	ARM	48	K. Hovhannesyan and T. Sarkisyan
Ashkenazi Jewish	*unknown*	AJW	9	Coriell (HD22)
Basque 1	France	BSQ	24	HGDP-CEPH
Basque 2	Spain		50	S. Alonso
Bulgarian	Bulgaria	BGR	616	G. Kirov and D. Levinson
Czechoslovakian	Czech Republic	CZE	10	Coriell (HD29)
English	UK (Cardiff)	GBR	184	G. Kirov and D. Levinson
Finnish	Finland	FIN	38	A. de la Chapelle and R. Kittles
French	France	FRA	29	HGDP-CEPH
German 1	Germany (Bonn)	DEU	149	D. Wildenauer and D. Levinson
German 2	Germany (Hanover)		30	R. Deka
Greek	Greece	GRC	8	Coriell (HD16)
Hungarian	Hungary	HUN	10	Coriell (HD26)
Icelandic	Iceland	ISL	12	Coriell (HD30)
Italian 1	Italy (Bergamo)	BGM	14	HGDP-CEPH
Italian 2	Italy (Sardinia)	SRD	28	HGDP-CEPH
Italian 3	Italy (Tuscany)	TUS	8	HGDP-CEPH
Orcadian	Orkney Islands (UK)	ORC	16	HGDP-CEPH
Russian 1	Russia (Novosibirsk)	RUS	25	HGDP-CEPH
Russian 2	Russia		51	T. Schurr
Spanish 1	Spain (Cantabria)	CTB	30	F. Leyva-Cobian
Spanish 2	Spain (Valencia)	VAL	56	E. Parra
Svan	Georgia	GSV	24	F. Leyva-Cobian
Kazakh	Uzbekistan	KAZ	40	L. Quintana-Murci
Turkmen	Uzbekistan	TKM	38	L. Quintana-Murci
Uzbek	Uzbekistan	UBK	37	L. Quintana-Murci
Mozabite	Algeria (Mzab)	MZB	30	HGDP-CEPH
Sahrawi	Western Sahara	WSH	28	F. Leyva-Cobian

Note.—HGDP-CEPH stands for the Human Genome Diversity Project, samples from the *Centre d'Etude du Polymorphisme Humain* (Paris, France)

Individual samples

The 180 individuals selected to be genotyped on the Illumina 317k arrays are a subset of the individuals sampled during the spring 2006 sampling tour of Europe described in Chapter 2. These 180 individuals are all female so that sex would not be a variable to control for when performing gene-phenotype associations; phenotypes measured for those persons include skin and hair pigmentation Melanin (M)-index, and iris luminance as described earlier (Chapter 2). Three-dimensional photographs of the face were also taken for future research on the underlying genetic basis of facial features. See Appendix E for ancestry and phenotype details.

Table 4.2 – 180 European individuals, genotyped on the Illumina 317k mapping array

Ethnicity (% based on 4 grandparents)	N *	Sampling Site †
Irish	**38**	**Dublin, Ireland**
Italian	**39**	**Rome, Italy**
Polish	**41**	**Warsaw, Poland**
Portuguese	**40**	**Porto, Portugal**
Scottish (75%) + Irish	a	Dublin, Ireland
French	b	Dublin, Ireland
Irish + English	c	Dublin, Ireland
English (75%) + Irish	d	Dublin, Ireland
Irish (75%) + Hungarian	e	Dublin, Ireland
Russian	f	Dublin, Ireland
German	g	Dublin, Ireland
Polish (50%) + French + Belgian	h	Rome, Italy
Italian (75%) + English	i	Rome, Italy
Italian (75%) + French	j	Rome, Italy
German (75%) + Dutch	k	Rome, Italy
Dutch + Italian	l	Rome, Italy
Italian (75%) + Polish	m	Rome, Italy
Bosnian + Montenegrin	n	Warsaw, Poland
Greek (50%) + Polish + German + Lithuanian	o	Warsaw, Poland
Polish (75%) + Lithuanian	p	Warsaw, Poland
Polish (50%) + German + Lithuanian	q	Warsaw, Poland
Portuguese (Angola?)	r	Porto, Portugal
Portuguese (Mozambique?)	s	Porto, Portugal
Portuguese (75%) + French	t	Porto, Portugal
Portuguese (75%) + Indian + African (AF=20% ‡)	u	Porto, Portugal
Italian Brazilian (IA=17% ‡)	v	Porto, Portugal

* Sample size or reference letter for single individuals
† All individuals were collected in the spring of 2006 by M.D. Shriver, M. Bauchet, B. McEvoy and local collaborators
‡ Partial AncestryByDNA 2.5 results for individuals with less than 80% EU ancestry component.

These 180 women were selected in sets of 45 from each of the four sampled cities in Europe, Dublin (Ireland), Rome (Italy), Warsaw (Poland) and Porto (Portugal). Approximately 40 of the women in each sample are autochthonous from the region where they were sampled and the rest are from different parts of Europe; the samplings were usually performed in Universities or scientific institutes, hence the presence of foreign individuals (Table 4.2). The Irish women are mostly from Dublin and surroundings, with only a few from western and northern counties. Italians are mostly from Rome and the surrounding regions of the center of Italy, but a substantial number are from the north, south and a few individuals have Sicilian and Sardinian ancestry (Figure 4.3B). Polish individuals were living in Warsaw but came from various parts of the country; German and Lithuanian ancestry is present in a few individuals (Table 4.2). Portuguese people mostly come from Porto, with a few from

the north and south. Some also had Portuguese ancestors from former African colonies of Portugal (Table 4.2).

Missing data were minimal; over the 317,503 available SNPs, 26 were eliminated for lack of genotypes and another 139 SNPs for being fixed for one allele in all samples. Average missing data for the remaining 317,477 SNPs was 0.86% per SNP. The four samples (N=45 each) have on average between 98.7% and 99.6% individual genotype per SNP (SD ranging from 0.7 to 1.4), and never less than 22 genotypes per SNPs.

EuroAIM selection

First, EuroAIMs were selected from the pool analysis by calculating the overall F_{ST} and keeping the highest ranking SNPs. F_{ST} was calculated by Weir's formula correcting for sample size[77,101] for the whole set of pools, and various F_{ST} thresholds resulted in different AIM sets. First, FST>0.01 among the 21 European pools (Figure 4.2D) yields 145,745 AIMs (Figure 4.5A and 4.5B). Applying a higher threshold (FST>0.03) reduces the AIM panel to 20,823 (Available at http://www.anthro.psu.edu/biolab/EuroAIMs.20k.xls), which can be reduced further to 12,035 AIMs by removing SNPs with null F_{ST} among the 4 groups—a manageable marker panel for the program *structure*. Note that this last AIM panel is specific to the 4 groups represented by individual genotypes (Irish, Italian, Polish and Portuguese).

In order to evaluate the informativeness of specific SNPs their F_{ST} can be compared to the empirical distributions of F_{ST} based on the full 317k SNPs, thus avoiding any model-based assumption on expected F_{ST} values[77]. To generate a simple empirical P-value I sort the SNPs by decreasing FST, and divide the rank of the SNP of interest by the total number of available SNPs (~317k). This technique is used here to give a rough estimate of the significance of F_{ST} values obtained for specific SNPs in candidate genes of interest (Table 4.4).

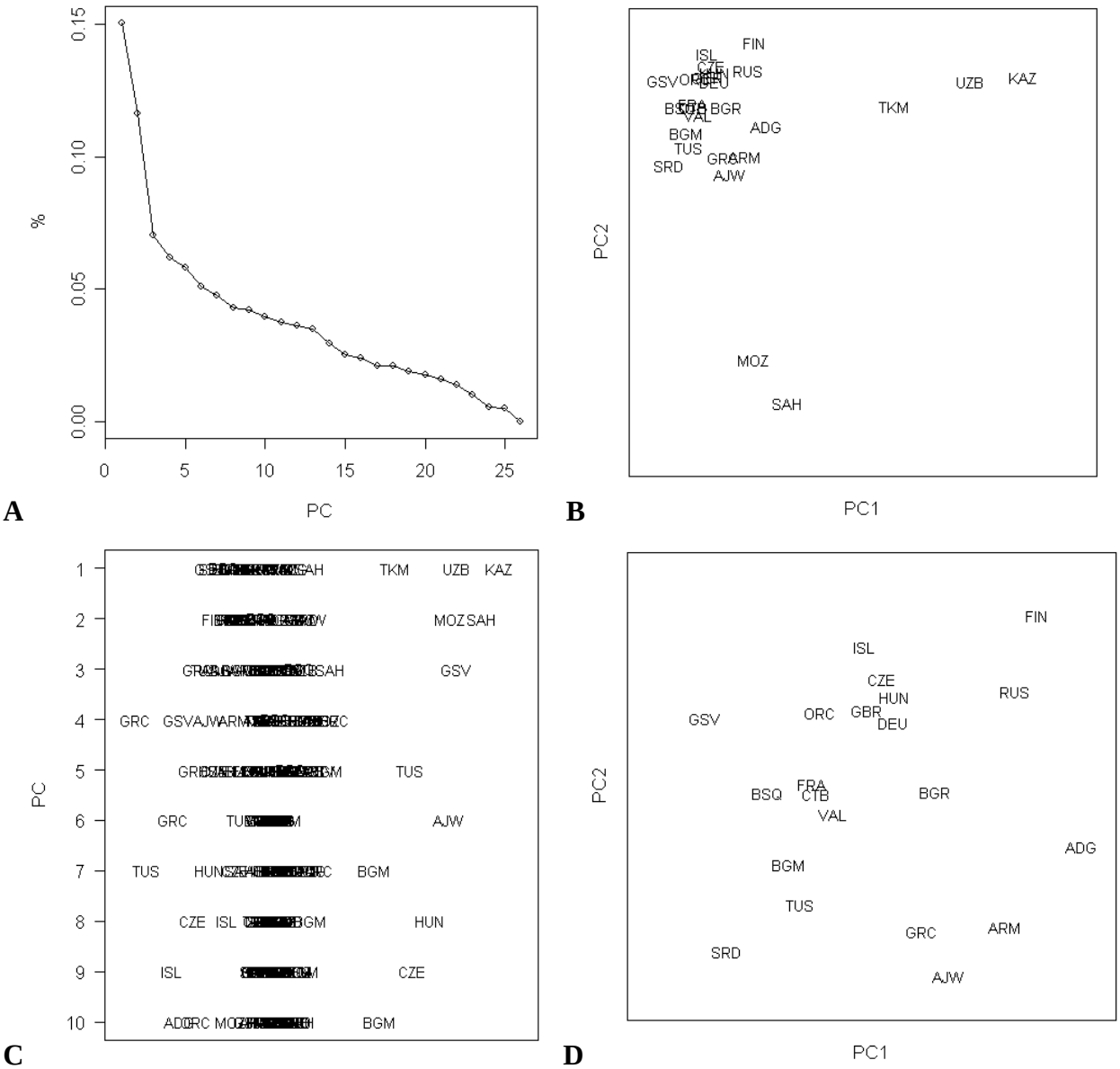


Figure 4.2 – PCA genotype data of 26 *pooled* population samples from Europe, Central Asia and North Africa. See Table 4.1 for 3-letter codes, sample sizes and provenance details. Each of the two German, Russian and Basque pools are merged. ***A,*** Percentage of the variance explained by each PC, ***B,*** Top two PCs reflecting geographic distribution, ***C,*** Top 10 PCs, ***D,*** Zooming in from plot B on the cluster of European pools.

Sample Pools

In PCA of all sample pools, the first two PCs are the most important, explaining respectively 15% and 12% of the variance; after PC3 (7%) the percentage of the variance explained decays steadily (Figure 4.2A). The first PC determines an axis of ancestry between Europe and Central Asia, whereas PC2 is marked by the separation between the European and North African samples (Figure 4.2B). The Georgian Svans drive PC3 and to a smaller extent PC4. Population pools with very small sample sizes, such as Coriell European and the two HGDP-CEPH Italian samples (Table 4.1), tend to drive the next

PC axes because of inaccurate frequencies (Figure 4.2C). If PCA analysis is performed with the European pools alone (e.g. without the five non-European samples) this effect is emphasized and the small pools drive all PCs (not shown), masking the European pattern observed in PC1 and PC2 in the context of the non-European populations (Figure 4.2D).

TKM and RUS have the highest mean heterozygosity (both ~0.351), whereas GSV has the lowest (~0.327) and all other pools were between 0.335 and 0.343.

Individual Genotypes

The multivariate analysis of the 180 Europeans is presented in Figure 4.3. The percentage of the variance explained is highest for the first three PCs, and stabilizes at PC4 and beyond. However, both the SKT and the ANOVA tests are highly significant for the first four PCs. PC1 mainly divides the northern samples (Polish and Irish) and the southern (Italian and Portuguese), and the SKT ($P<0.0001$) and ANOVA test both show very significant correlation (ANOVA $R^2 = 0.98$). PC2 (not shown) is mostly driven by the Irish heterogeneity and a slight separation from other populations, and PC3 (not shown) separates the westernmost samples (Irish and Portuguese) from the easternmost (Italian and Polish). Both PC2 and PC3 have however relatively low correlations on the ANOVA test (respectively R^2=0.49 and 0.74); PC2 slightly separates the Irish from all other individuals, in a way that scatters them throughout nearly the entire range of PC2 values (not shown). PC3 offers a weak separation between eastern and western groups, but with many scattered individuals (not shown). These "noisy" patterns may be due to the high number of low-differentiated markers across the four groups (nearly half have null overall F_{ST} among the four groups). Like conventional PCA, PCoA extracts trends of highest variabilities, irrespective of whether they provide clustering or not; "this may give rise to the situation where the chosen principal component corresponds to the attribute with the highest variability but without any discriminating power"[102]. Sophisticated supervised learning algorithms can be applied to extract features[102], but with a simple combination of PC2 and PC3 together the ANOVA correlation becomes 0.93 which is very near that of PC1 ($R^2 = 0.98$); therefore, this combination was chosen as a better representation of the next significant trend among genetic distances (Figure 4.3C).

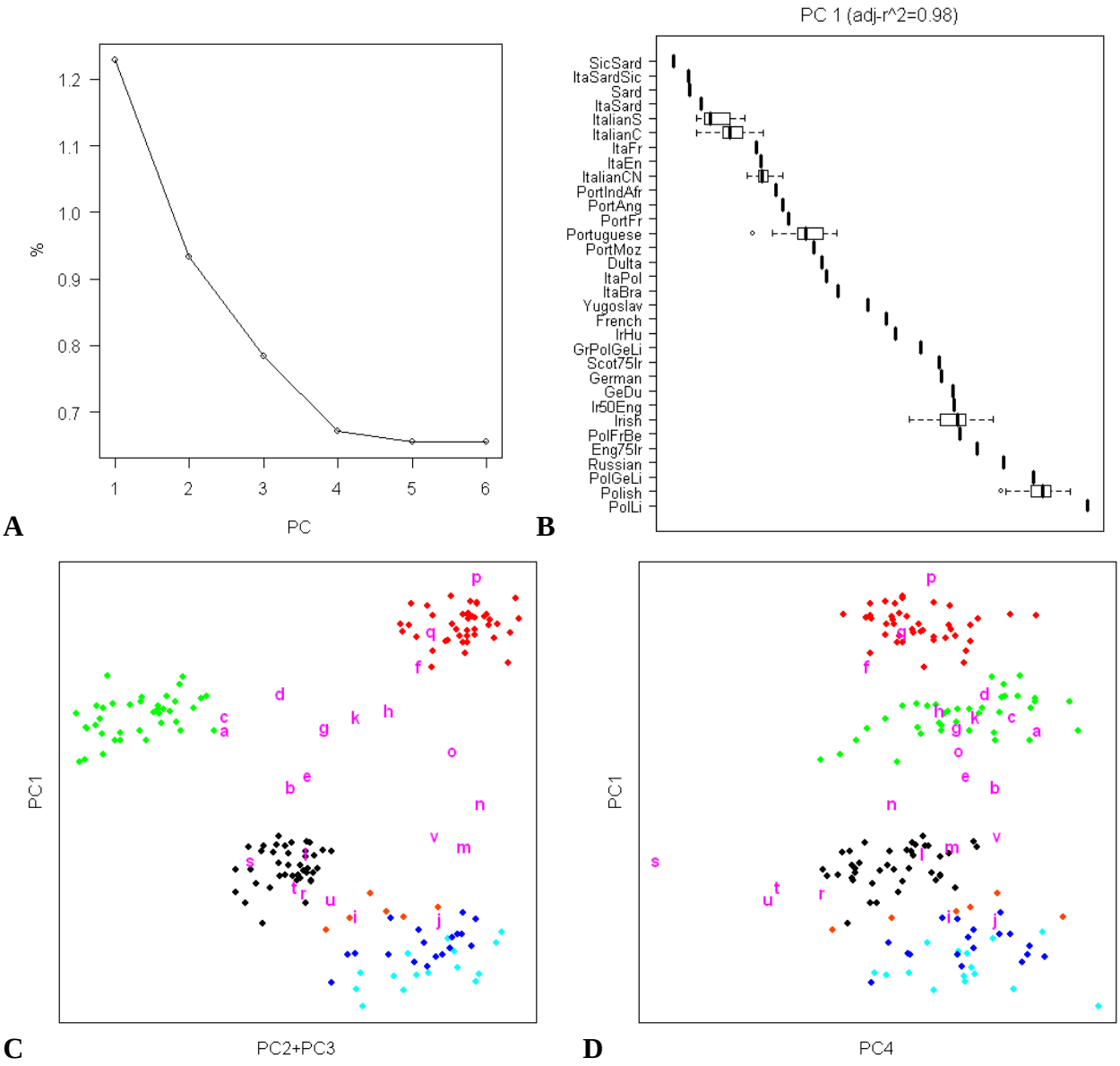


Figure 4.3 – PCoA of all 180 *individuals* (Table 4.2), using the full 317k SNP panel. ***A,*** Decay of the percentage of the variance explained by top 10 PCs. ***B,*** PC1 with individuals of specific ancestry. Abbreviations correspond to ethnicities listed in Table 4.2 except for Italians sub-groups (ItalianS=South, ItalianC=Center, ItalianCN=Center-North) and island individuals (Sard=Sardinian, Sic=Sicilian). ***C,*** Top 3 PCs combined. Letters refer to individuals listed in Table 4.2. Irish are in green, Polish in red, Portuguese in black, ItalianCN in orange, ItalianC in blue, ItalianS in cyan. ***D,*** PC1 with non-European axis PC4.

The Irish, Italian, Polish and Portuguese samples have similar heterozygosity (Figure 4.4) in agreement with their similar cluster densities in PCoA (Figure 4.3). In the SKT (10 P-values) over all strictly Irish, Italian, Polish and Portuguese individuals (respectively N=38, 41, 41 and 40), only the Polish group yields no significant internal structure. The group with the highest heterozygosity, the Italian, is the only one that could meaningfully be subdivided into 3 groups created from declared grandparental ancestry (Figure 4.3C), which are significantly different (ANOVA). The Irish had some slight

separation (not shown) between the Leinster (Dublin and surrounding counties) and the rest, but the ANOVA significance remains poor. The Portuguese showed no specific pattern depending on individual origins within Portugal, but the sample was mostly composed of individuals around Porto or admixed from different regions.

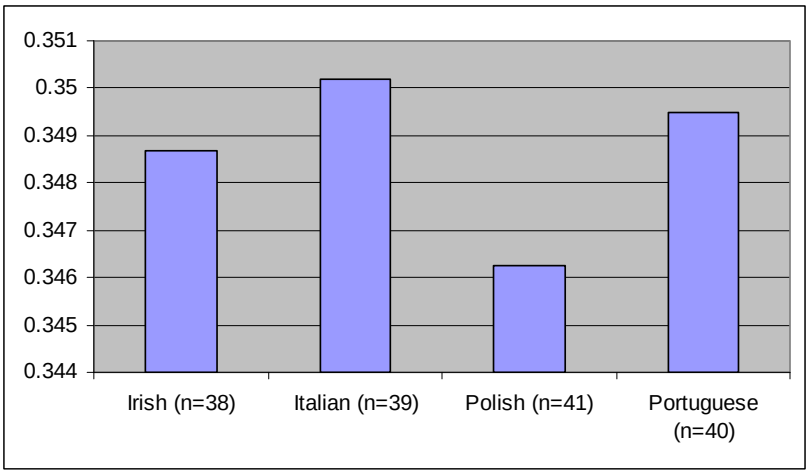


Figure 4.4 – Average heterozygosity for the four samples (leaving out foreign and admixed individuals, Table 4.2).

EuroAIMs

A first set of 145,745 AIMs were selected for having FST>0.01 among the European pools. PCoA using the 180 individuals and this AIM panel (Figure 4.5A and 4.5B) preserves well the pattern of structure observed with all 317k SNPs (Figure 4.3C and 4.3D), as well as the PC significance. In order to find a more manageable marker panel to be tested in *structure*, I further reduced the number of EuroAIMs by setting the F_{ST} threshold at 0.03, which yielded 20,823 EuroAIMs—again based on the pooled samples PCA. I then eliminated SNPs that had zero F_{ST} among the four main groupings (Irish, Italian, Polish and Portuguese). This resulted in a set of 12,035 AIMs, which also preserve the overall pattern (Figure 4.5C), even though the significant PCs have slightly different meanings. PC1 no longer distinguishes the Irish from the Polish, but rather clusters the Irish and Polish together, away from the Italian; with the Portuguese still positioned in between, and the Italian north-south pattern still present. PC2 and PC3 reconstitute the gap between the Irish and the Polish. PC4 does not represent the non-European admixture anymore, and neither do any of the other top 6 PCs, which is not surprising since this panel of 12,035 AIMs is now very European oriented.

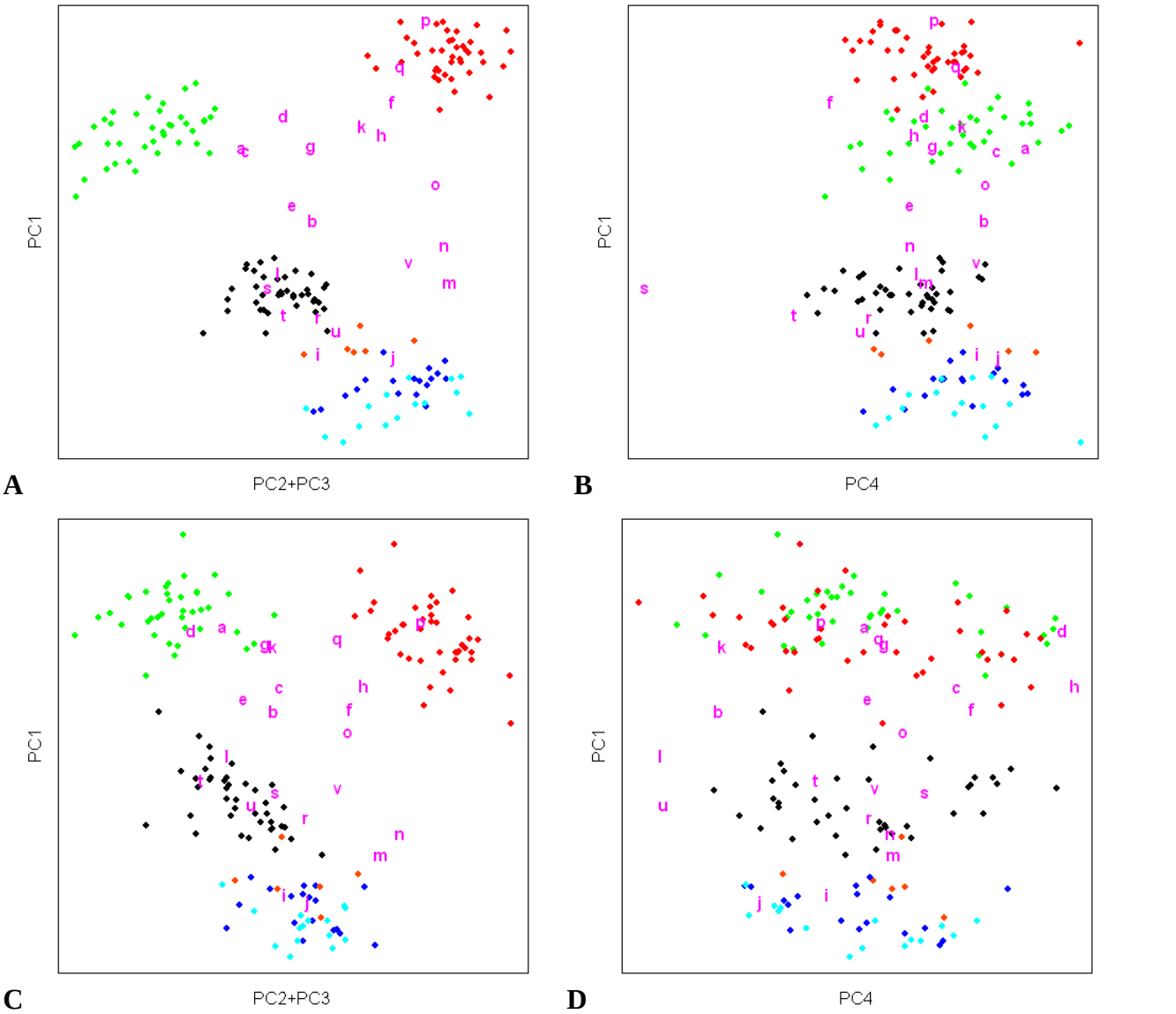


Figure 4.5 – PCoA of all 180 *individuals* (Table 4.2), using AIM panels. Abbreviations correspond to ethnicities listed in Table 4.2 except for Italians sub-groups (ItalianS=South, ItalianC=Center, ItalianCN=Center-North) and island individuals (Sard=Sardinian, Sic=Sicilian). Irish are in green, Polish in red, Portuguese in black, ItalianCN in orange, ItalianC in blue, ItalianS in cyan. ***A and B,*** PCoA with 145,745 AIMs of F_{ST}>0.01 among 21 European pools (Figure 4.2D). ***C and D,*** Using SNPs with F_{ST}>0.03 among the pools and keeping the 12,035 SNPs of null F_{ST} among the 4 groups (AIM panel used in Figure 4.6).

Considering that the 12k SNP set represents well the overall European structure among the 180 individuals, I used it to run *structure* on all 180 individual genotypes (Figure 4.6). Clusteredness (G) is highest for K=2 (G=67%), below 50% for K=3 and 4, and high again for K=5 (G=57%). Posterior probability is highest for K=4 (PP=1).

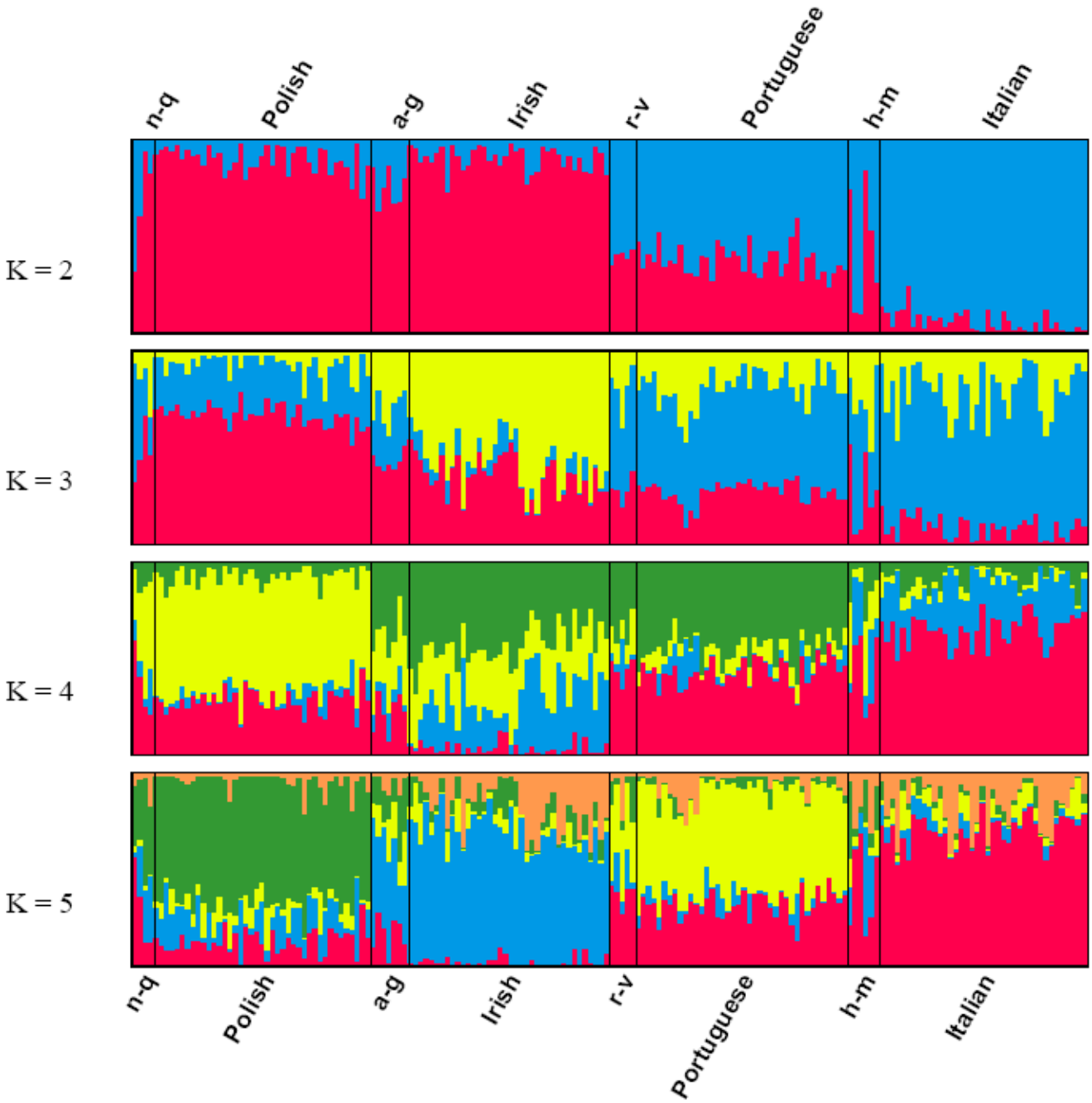


Figure 4.6 – Bayesian analysis with *structure* using ~12k AIMs (Figure 4.4C). Country labels represent autochthonous individuals from the four corresponding sampling sites. Foreign individuals are represented by letters a-v in the same order as in Table 4.2.

Phenotypes

Eye, hair and skin pigmentation phenotypes were available for most of the 180 women (Appendix E). Hair pigmentation is the sparsest dataset because individuals with dyed hair were not sampled. Those three phenotypes were also measured in many other individuals from these sampling sites and from other sampling sessions in France and the US (Chapter 2); it is therefore possible to compare the phenotypic correlations in all available individuals to the present subset of 180 women.

Table 4.3 – Phenotype inter-correlations of Irish, Italian, Polish and Portuguese samples

	All individuals			**Persons typed on 317k**		
	R^2	P	N	R^2	P	N
Skin v. Eyes	-0.38	**<0.001**	534	-0.41	**<0.001**	159
Irish	-0.14	0.086	144	-0.28	0.095	37
Italian	-0.16	0.073	123	-0.36	0.019	41
Polish	-0.028	NS	83	0.034	NS	41
Portuguese	-0.072	NS	184	-0.11	NS	40
Hair v. Eyes	-0.5	**<0.001**	417	-0.4	**<0.001**	124
Irish	-0.14	NS	121	-0.018	NS	30
Italian	-0.31	**<0.01**	80	-0.16	NS	27
Polish	-0.015	NS	62	-0.051	NS	33
Portuguese	-0.34	**<0.001**	154	-0.16	NS	34
Skin v. Hair	0.37	**<0.001**	425	0.37	**<0.001**	124
Irish	0.0039	NS	123	0.033	NS	30
Italian	0.31	**<0.01**	85	0.24	NS	27
Polish	0.32	**<0.01**	63	0.38	0.028	33
Portuguese	0.12	NS	154	0.049	NS	34

Note.—The first line of each phenotype comparison uses all individuals together. P values in bold are significant at the 0.01 or 0.001 level. NS means not significant at the 0.1 level.

Since the three pigmentation phenotypes vary substantially across Europe (Chapter 2) correlations between each phenotype are high when all 180 individuals are used. However these correlations mostly disappear within each of the four main country groups (Table 4.3, "Persons typed on 317k"). To some extent this might be an artifact of the smaller group size because when all available individuals are used some within-group correlations become significant (Table 4.3, "All individuals")

Using the full set of individuals sampled for the four populations I also verified that, for each of the 12 phenotype-group combinations (three phenotypes, 4 groups), no significant difference (Wilcoxon P>0.1) could be detected between male and female distributions—each of the eight cohorts had sample sizes ranging from 25 (male Polish hair) to 123 (female Portuguese eyes).

As shown in Chapter 2, eye color is highly bimodal whereas hair and skin approximate normal distributions (Figure 4.7). Most Irish and Polish are in the light-eye cluster, although about one quarter

of the Polish are with most Italian and Portuguese in the dark-eye cluster (Figure 4.7B and 4.7C). Also, both the reduced groups of the 180 individuals and the larger groups of all available individuals show that eye color is the strongest determinant of north-south ancestry, followed by hair color. Skin color is only different in the Irish individuals (Figure 4.7A and 4.7B), which is in agreement with the findings presented in Chapter 3.

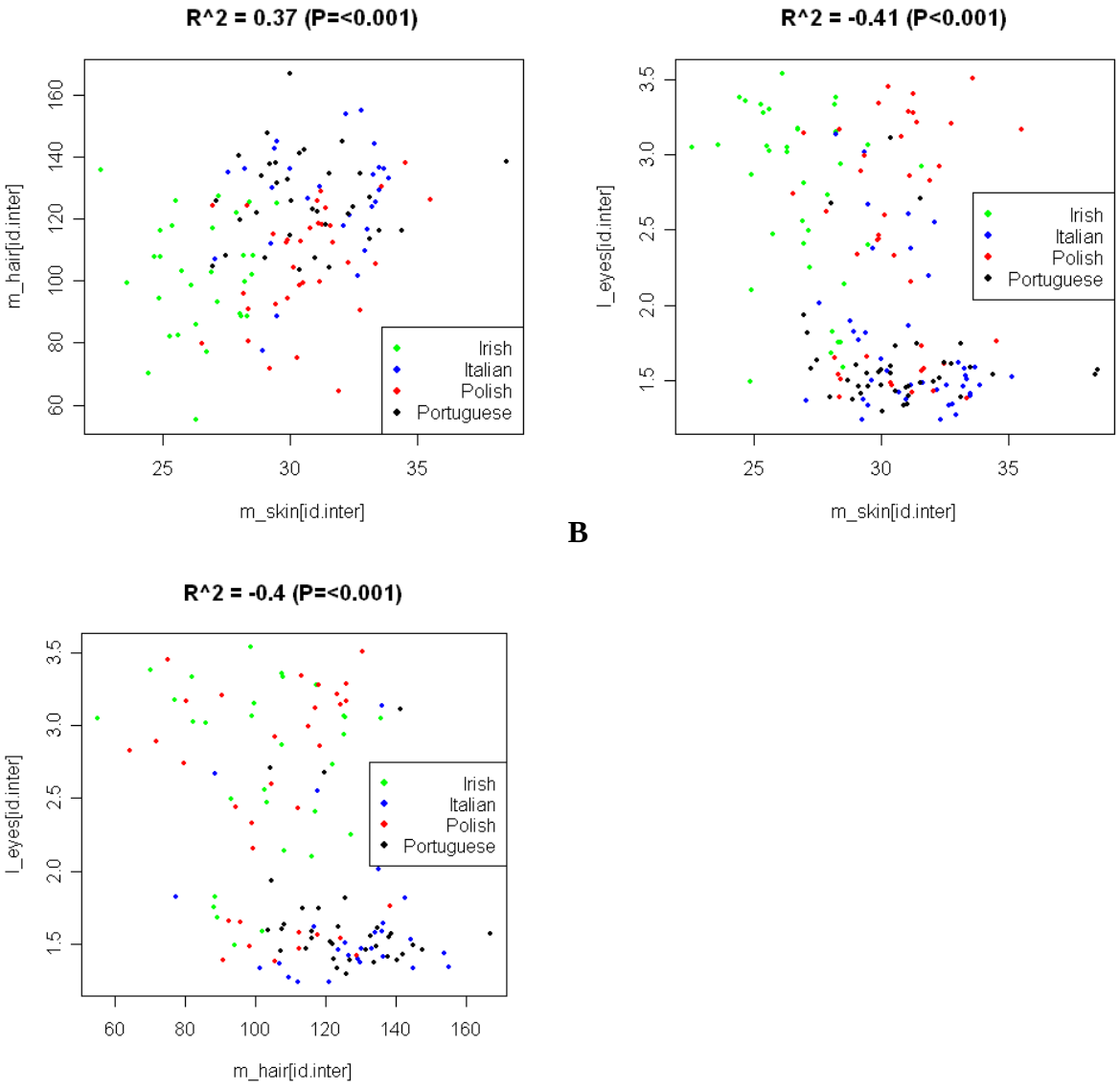


Figure 4.7 – Phenotype inter-correlations.

Functional Genes

Pigmentation candidate genes[103] as well as *LCT* (a gene known to cause stratification in Europe[37]) were examined in particular for their variation among the four groups. The *DTNB* gene was added because it is on the same pathway as *DTNBP1* and contains SNPs that were in the best EuroAIMs selected in the 10k analysis (Chapter 3). For each gene, SNPs with pairwise $F_{ST} > 0.1$ (.i.e. between each pair of groups) were retained (Table 4.4). For each SNP an empirical P-value was calculated by comparing its F_{ST} value to the empirical F_{ST} distribution (See Methods, EuroAIMs Selection).

Some pigmentation genes did not have SNPs of high F_{ST} among the four groups, possibly because they had few SNPs on the 317k mapping array to begin with (*ASIP, MC1R, MATP, EDN3, ADAM17, KITLG, TYR, DCT, MITF, EDAR, LYST*). However, despite only having 5 to 7 SNPs that could be tested, *TYRP1*, *SLC24A5* and *LCT* had SNPs of interest (Table 4.4).

Table 4.4 – Pigmentation candidate genes and SNPs with highest pairwise F_{ST} among the 4 European groups (individuals typed on the 317k mapping arrays)

		Overall Weir F_{ST}	Pairwise Weir F_{ST}					
dbSNP ID	Gene		Irish-Italian	Irish-Polish	Irish-Portuguese	Italian-Polish	Italian-Portuguese	Polish-Portuguese
rs2322659	*LCT*	0.206 ***	0.506 ***	0.238 ***	0.263 ***	0.101 .		
rs2874874	*LCT*	0.065 **	0.205 ***		0.123 **			
rs3754689	*LCT*	0.097 ***	0.277 ***	0.104 *	0.110 **			
rs3754690	*LCT*	0.057 *	0.187 ***	0.105 *	0.136 **			
rs4954633	*LCT*	0.048 .	0.161 **					
rs4236167	*DTNBP1*	0.041 .			0.102 *			0.112 *
rs9396592	*DTNBP1*	0.039 .			0.129 **			
rs1465541	*DTNB*	0.051 .	0.120 *		0.125 **			
rs989869	*OCA2*	0.048 .				0.132 **		0.109 *
rs3947367	*OCA2*	0.041 .				0.116 *		0.101 *
rs1004611	*OCA2*	0.049 .				0.117 *		0.118 **
rs3794604	*OCA2*	0.065 **	0.153 **		0.148 **			
rs4778232	*OCA2*	0.133 ***	0.150 **		0.211 ***	0.165 **		0.228 ***
rs8024968	*OCA2*	0.064 **	0.143 **		0.138 **			
rs7170869	*OCA2*	0.127 ***	0.2 ***		0.223 ***	0.144 **		0.165 ***
rs1597196	*OCA2*	0.068 **	0.101 *		0.120 **			0.105 *
rs1448485	*OCA2*	0.081 **	0.159 **		0.154 **			
rs7496968	*OCA2*	0.047 .						0.156 **
rs7179994	*OCA2*	0.076 **						0.202 ***
rs7495174	*OCA2*	0.119 ***	0.216 ***		0.199 ***	0.146 **		0.129 **
rs2433356	*SLC24A5*	0.068 **			0.110 *			
rs2675348	*SLC24A5*	0.132 ***		0.237 ***	0.242 ***	0.119 *	0.122 **	
rs3743288	*SLC24A5*	0.07 **		0.111 **	0.134 **		0.102 **	
rs2075508	*TYRP1*	0.051 .			0.11 **			

Note.— Empirical P-values are noted as: . when <0.05, * when <0.01, ** when <0.005, *** when <0.001. Cells are left blank when the condition of pairwise $F_{ST} > 0.1$ is not met.

Discussion

Population pools

The separation of Uzbekistan, Turkmenistan and Kazakhstan populations from European and African pools (Figure 4.2C and 4.2D), which is the main trend of the pool analysis, is in accordance with geographical orientation of these respective Central Asian countries from southwest to northeast (Figure 4.1B). This is consistent with the route through which agriculture and farming spread from the Near East during the Neolithic revolution, mirroring similar events in Europe[25]. The two North African populations drive the second main trend of the pool multivariate analysis, which also shows a trend consistent with their geographical distance from Europe. Interestingly, despite their more western location, the SAH sample appears east of the MZB and European pools; in fact this happens to be consistent with the putative Yemeni Arab heritage of this Western Saharan population. Individual genotyping of the PC1 and PC2 AIMs on North African and West Asian population samples would be a starting point to study the genetic structure in these regions.

Within Europe, the Georgian Svan (GSV) pool stands out from other Europeans in PCA pool analysis in a peculiar way (Figure 4.2). At first sight, knowing that this sample represents a small isolated mountain population, this pattern may be interpreted as the result of drift, and therefore this pool could be discarded in the search for AIMs for not representing a major contribution to current European variation. However, despite their small population size (~15,000), the Svan have retained high levels of genetic diversity in HLA genes[104]. In practice GSV has by far the lowest heterozygosity of all pools, which may explain their high degree of differentiation—but not the direction of this separation. Second, in total contradiction with geographical expectation they are at the opposite end of the other Caucasus population represented here, the Adygei (ADG, Figure 4.2D), geographically located just north of Georgia, near the Black Sea. Some genetic separation is actually expected considering that the Adygei speak a North Caucasian language, whereas Svanetian, as the other Georgian languages, belong to the Kartvelian family. Third, the GSV pool is closest to the Basque—even more so when the two Basque pools are not merged (not shown). Interestingly, both the Basque and the Georgians are commonly considered linguistic isolates; despite recent studies suggesting absence of genetic ties between the two ethnicities, some have found them linked by linguistic and toponymic similarities[104]. However, no similarity was observed between Georgians and Basques using mtDNA analysis; haplogroup H emerged and diversified between 20,000 and 10,000 years ago from or near the Basque region but is absent in the Caucasus[105]. This may be an artifact of the use of a single uniparental marker, or that Basques and Georgians have common roots from long before the Neolithic.

The spatial distribution of the European pooled data is also consistent with the patterns of European genetic variation we observed using DNAPrint's EuroAIMs (Chapter 2) and from the 10k analysis (Chapter 3). Finns are still highly separated on the main axis (Figure 4.2B and 4.2D), on the opposite pole of a cluster of southeastern European and Near Eastern populations (Greek, Armenians, and Ashkenazi Jewish), like in the analysis of individuals with 10k mapping arrays (Chapter 3). The present analysis also includes a Sardinian population pool which also clusters "South" on PC2 but is only second to the Georgian Svans in being the most distant genetically from the Asian pole of ancestry (Figure 4.2B and 4.2D). Although we do not have further details on this particular Sardinian sample, Sardinia was peopled before the Neolithic and throughout the ages experienced both drift and population growth. Sardinians are also known from analyses of classic markers to be an outlier from the rest of Italy[11]; this is confirmed here. Additionally, their location in PCA space also suggests gene flow from the Middle East or North Africa (Figure 4.2B and 4.2D).

The Basque pools (combined in PCA analysis) appear again very separate from the Mediterranean Spanish, as in the 10k mapping array analysis (Chapter 3). The present analysis allows placing them in the context of a French sample and another Spanish sample from Cantabria, a region located west of the Basque country. In opposition to geographical expectation (Figure 4.1A), the French and Cantabrian pools appear very near to one another, and away from the Basque.

In the northern European cluster, a reverse geographical pattern can also be observed, with the Czech and Hungarian appearing "North" (i.e. higher on PC2) of the German pools, oriented in the direction of the Icelandic pool. The island of Iceland was settled in the 9th century AD mainly by Norse populations, and minor gene flow from the British Isles may have also happened; previous analysis of classic genetic markers are consistent with this history[11]. In agreement, the Icelandic pool appears as the northernmost population in this analysis, only second to the Finns who are, additionally, strongly separated from other European populations in direction of Central Asian populations (Figure 4.2B). Orkney Islanders appear slightly separated from English and German pools, which is not surprising for an island population in relative isolation from its population of origin on the continent. In general concordance with geography, the Russian pool is located "East" (i.e. high on PC1) of the Germanic populations and intermediate between the Finns and the Adygei population of the Caucasus (PC2, Figure 4.2D). At this point further evidence is needed to interpret the intriguing pattern of the Czech and the Hungarians; they appear north of the Germans, rather than east as would be expected by their geography. This would be possible with lager samples of known provenance and individuals genotyping.

Individual genotyping

The main pattern of the 180 individuals genotyped for 317k SNPs appears to be the separation between the two northern samples (Irish and Polish) from the two southern (Italian and Portuguese). However, the Portuguese are intermediate between the Italian and the Irish in PC space (Figure 4.3C), which is in contradiction with the geographical north-south alignment of their respective countries (Figure 4.1A). This is also confirmed by the F_{ST} calculation, which shows that the closest to the Irish are the Portuguese, followed by the Polish, and the Italian (Table 4.5, first column). The affinity between the Portuguese and Irish samples, higher than would be expected based on geography, is consistent with a previous study of mtDNA evidence showing connections among peoples along the Atlantic coasts of Europe, dating back to the end of the last Ice Age[93].

Table 4.5 –Mean F_{ST} x10^3, using all 317k SNPs

	Irish	Italian	Polish
Italian	10.6		
Polish	9.0	12.0	
Portuguese	8.8	7.6	10.2

The second most important pattern (PC2) is mostly driven by an apparent Irish heterogeneity and slight separation from other populations (not shown), which somewhat parallels the pattern of skin color differences (Chapter 3). However the Irish are far from having the highest heterozygosity, so it is unclear what this pattern may represent. In fact, the Irish become as clustered as the other groups when PC2 is combined with PC3 to form a more significant component than either on its own (Figure 4.3C). When added, PC2 and PC3 also explain more of the variance than PC1 (Figure 4.3A). Therefore the east-west trend is overall at least as important, if not more, as the north-south axis in Figure 4.3C. The Atlantic connection between Portugal and Ireland is also further emphasized, as well as the difference between Italians and Poles. Italian individuals in this sample however are mostly from the center and the south of Italy. The few individuals with northern Italian ancestry do separate significantly from other Italians on PC1, not only "north" but in the direction of the Portuguese individuals (Figure 4.3B), possibly due to gene flow along the Mediterranean coast into the Iberian peninsula. Interestingly, southern Italians also separate significantly from individuals of central Italy (Figure 4.3B and 4.3C), the southernmost appearing to be the individuals with Sardinian and Sicilian ancestry. One possible hypothesis for this latter pattern could be a relative degree of African affinity, as was suggested above by the position of the Sardinian pool relative to North African pools (Figure 4.2B). However, the individuals of Sardinian and Sicilian ancestry show no African affinity; in fact, these southernmost

Italians on PC1 are on the opposite end of PC4 from the individuals of African ancestry (Figure 4.3D). The last significant trend is driven by only a few individuals of known or suspected admixture from former Portuguese colonies of Africa (individuals; r, s, t and u) or the Americas (individual v). Both Sicily and Sardinia are however relatively diverse and the few individuals in this sample should be augmented to investigate these patterns, in the context of individuals from North Africa proper.

An important aspect of the overall structure pattern is that the European individuals who don't belong to any of the four groups (a-o, q and t) are spatially distributed within the PC space defined by the four groups. The most extreme individuals tend to also have the most outlying geographical ranges (e.g. the Sicilian-Sardinian individual and the Pole with Lithuanian admixture are at the far ends of PC1). Among the individuals suspected of having African admixture (s, t, u and v), only one (u) showed a significant AF component on DNAPrint AncestryByDNA 2.5 test. These individuals however are most likely to have admixture from Angola or Mozambique, in southern Africa—very different in mtDNA composition[106]. This is consistent with the possibility that AncestryByDNA 2.5 is highly specific to detecting western African ancestry; lack of AF ("Sub-Saharan African") ancestry in an individual therefore is inconclusive in terms of ancestry for other parts of Africa.

The Bayesian analysis with *structure* confirms the PCoA conclusions (Figure 4.6). Additionally, although this representation lacks the sense of spatial distribution of PCoA, it is important to note that the *structure* clustering pattern underlines inter-population admixture and gene flow. Indeed, even though individuals in each of the four groupings consistently have similar profiles across runs, and can be assigned to a category based on this profile, in particular for K=5, the within sample clusteredness remains low. This is consistent with the relatively high amounts of gene flow throughout Europe since antiquity.

The SKT detects significant internal stratification (combined-P<0.001 on first two PCs) within the three samples of highest heterozygosity namely the Irish, Italian and Portuguese. The structure within the Italian sample can be laid out as three subgroups (Figure 4.3B and 4.3C) distributed along PC1; the Center-North, Center and South Italians (respectively labeled ItalianCN, ItalianC and ItalianS). This is not surprising in light of the separation among the few Italians presented in Chapter 3, as well as previous work covering the country more broadly and using between 31 to 319 classic gene frequencies, which showed complex PCA clines across the peninsula and Sicily[11]. Better delineation would therefore be possible with more samples from the northeast, northwest, the extreme south and the two main islands (Sicily and Sardinia). The structure within the Irish sample could not be disentangled. This is possibly due to the highly skewed sample in favor of Leinster and Dublin area, where individuals come from various parts of Ireland; hence this Irish sample contains many

individuals admixed from different regions and few can trace grandparental origins to a single region. It remains to be seen, with a sample individuals with better controlled ancestry, how this stratification would be geographically distributed. A similar situation was present in the Portuguese sample, where most individuals were from the Porto region. The Polish sample was the most nationally widespread of the four samples, and still presented no detectible structure and the lowest heterozygosity. Considering the PCoA position and *structure* proportions of the Russian individual (f), it is possible that this relative homogeneity extends north and east beyond the borders of Poland. The two individuals with German ancestry (g and k) however were quite separated from the Polish cluster in PCoA (Figure 4.3C), suggesting that the relative homogeneity of the Polish cluster doesn't extend south.

Phenotypes

In Chapter 2 we saw using larger samples of the four populations (Irish, Italian, Polish, Portuguese) which ones are significantly different from each other for different phenotypes (skin and hair pigmentation and eye luminance). Are the four female samples presented here a good representation of this variation? First, it seems that no major bias was introduced by using exclusively females, since for each of the four populations considered here the overall pigmentation distributions for females are not significantly different from males (see Results). The distributions within the 4 groups also appear relatively well preserved by using the reduced sets of ~40 individuals per population. However Table 4.3 shows that some correlations are lost compared to the full sets of available individuals, namely the correlation between hair and eyes (for Italian and Portuguese) and between skin and hair (for Italian and Polish). This may limit the power of association mapping if performed specifically on these few phenotypes and samples. This is not a concern if all four samples are used together, because the phenotypes are at least as strongly inter-correlated with all individuals and with the ~40 individuals per sample (full set correlations in Table 4.3).

Known and candidate pigmentation genes appear to hold SNPs that are also EuroAIMs. Five candidate pigmentation genes, *DTNB*, *DTNBP1*, *TYRP1*, *SLC24A5* and *OCA2*, stand out in having multiple high F_{ST} SNPs among the four Europeans groups, which confirms the previous suggestions that they are functionally related to skin lightening in Europeans[103,107]. Whereas most of these genes (*DTNB*, *DTNBP1*, *TYRP1* and *OCA2*) appear largely differentiated between each of the northern (Irish and Polish) and the southern (Italian and Portuguese) groups, *SLC24A5* also has high F_{ST} SNPs along the east-west axis (Irish-Polish and Italian-Portuguese, Table 4.4). The *OCA2* gene shows the most striking pattern with as many as 12 SNPs of high F_{ST} between at least one northern and one southern group (Table 4.4). This confirms the position of *OCA2* as a critical gene in determining eye color in

Europeans[68], including a SNP (rs7495174) recently found to be directly associated with blue eye color[108]. Finally the *LCT* gene, involved in human variation in lactose tolerance and shown to be under strong positive selection[109], holds SNPs mostly differentiated between the Irish and each of the other three groups (Table 4.4). With nearly 90% lactase tolerance, Ireland has one of the highest rates in Europe, with Scandinavia and Poland[109,110]. It remains to be seen why we observe striking differences between the Irish and the Polish at several *LCT* SNPs and whether, like in Africa, Europe also holds a case of convergent adaptation at this gene[111].

The 180 women genotyped for 317k SNPs hold great promise for studying gene-phenotype association (pigmentation and facial features), and SNP density at some genomic regions may be good enough to examine linkage disequilibrium patterns.

Chapter 5

Conclusion

"All models are wrong, some models are useful"—George Box[112]

Methods

The genetic distance based methods used here (PCA and PCoA on ASD matrices) proved especially useful to make sense of large marker sets such as the 10k and 317k mapping arrays. Performing Bayesian analysis (with *structure*) was found to be impractical for SNP sets larger than a few 10,000s of SNPs, because each run time for a specific K would take more than two days on the supercomputing Linux cluster Lion-XO of the High Performance Computing Group at Penn State. Using PCA and PCoA methods in parallel with Bayesian analysis helped validating the significance and biological importance of genetic structure pattern. PCoA and PCA methods could be even further improved by additional steps combining PCs to detect optimal clustering patterns, as shown by adding PC2 and PC3 in Chapter 4 (Figure 4.3) or any combination optimizing clustering[102]. For both methods, the interpretation of results can be done in several ways.

In Bayesian analysis, the posterior probability (PP) can be at odds with overall clusteredness (G). PP only indicates the model of highest likelihood, i.e. the most likely K number of clusters; but any interpretation should take into account that the model is highly dependent on which populations are represented by the individuals—I showed the importance of having a balanced representation from the putative groups (Chapter 2). On the other hand G gives a measure of the relative degree of stratification among different clustering models, hence which models may be useful to screen for AIMs—to eventually detect and control for stratification. Most important for concluding a pattern is robust is that it persists across runs, across slight variations of individuals and that it is preserved even when the marker set is reduced from the full panel (Chapter 3 and 4).

In PCoA, the portion of the variance explained by each PC (i.e. their relative Eigen value) measures the relative importance of a given PC, but not its statistical significance. Although these two concepts are related they do differ markedly. For instance in the 10k analysis of Eurasian and African populations, PC2 explains 10 times less variation compared to PC1, but it still is an axis of statistically and biologically significant clustering (Chapter 3, Figure 3.1B). Therefore sharp plunges of the relative Eigen value for a specific PC must be interpreted with caution as they do not necessarily tell which axes of stratification are significant—only their relative importance. The SKT and ANOVA test measure different types of PC significance. The ANOVA test measures the correlation between group labeling and a given PC, therefore measuring also the relevance of the group labels. The SKT on the other hand is independent of group labeling and measures how consistent a PC is throughout all chromosomes. A further improvement would be to perform the SKT with permutations among all unlinked SNPs rather than entire chromosomes, which should prove useful when fewer SNPs are

available. With increasingly denser SNP typing, the SKT could also be adapted to detect differences in ancestry orientation between specific genomic regions, an important part of admixture mapping[113].

Practical Considerations

The observation that variation within continents tends to be more continuous than between continents[4]—at least before recent world migrations, is verified for Europe. Indeed, in all previous chapters, from 313 to ~317k SNPs, African and Asian individuals (or pools) always map away by several orders of magnitude from a dense cluster of Europeans. This dense cluster of Europeans, however, was shown to be also structured, using independent populations and partly independent marker panels (the 10k and 317k arrays have less than 3,000 SNPs in common). When the number of AIMs is increased, groups of individuals tend to overlap less and to cluster more, showing that apparent absence of stratification was caused by too few markers or poor informativeness of the AIMs.

Recent studies have also demonstrated the presence of European stratification which can create false positives in gene association studies[36-38]. Although the intra-European levels of stratification are relatively low, as sample sizes (and number of samples) in association studies become larger in order to track down rare or complex traits, it becomes increasingly crucial to adjust for stratification. Also, as demonstrated by a recent paper[37], particular genes and phenotypes might be especially sensitive to the effects of population stratification in European-based samples. Indeed, this study shows spurious association between height and a SNP in the *LCT* gene[37]; these two traits follow a similar southeast-northwest trend as the three pigmentation trait, whose candidate genes have F_{ST} at least as high as *LCT* (Chapter 4). I have presented several EuroAIM panels in Chapters 3 and 4, which can help control for this stratification[39] for future investigations of the genetic basis of these traits.

Additionally, since women from the four groupings (Ireland, Italy, Poland and Portugal) cluster apart so clearly according to their country of origin, and because linkage disequilibrium (LD) is relatively low in Europeans[114], admixture mapping[115] may become an option for studying genes that contribute to complex traits and common diseases in Europeans. It would be a matter of finding admixed individuals between these populations, such as the few presented in Chapter 4; this should be easily possible in Europe as grandparental ancestry (and beyond) can be traced from written records. Also, migration patterns of last century include, for instance, Polish and Russian immigration in France and more recently in Ireland; Portuguese, Spanish and Italian immigration in France; Turkish immigration in Germany, etc. In view of the strong observed clustering, especially with the EuroAIMs

selected from the 317k arrays, it should be possible to find enough individuals for powerful admixture mapping in Europe (e.g. ~2,500 patients typed for ~2,500 markers)[113].

EuroAIMs can also be useful in forensics, as one more line of evidence for investigators trying to uncover identifying information on a person based on DNA[69,116]. For diseases that have higher prevalence in certain populations, genetic ancestry alone might be a useful indicator (e.g. Ashkenazi Jewish ancestry and risk of being a carrier of Tay Sachs). So far, little genetic disease susceptibility variation was found among European populations, but pigmentation phenotypes such as skin, hair and eyes color do vary across Europe (Chapter 2). These traits are strongly associated to risks of melanoma and non-melanoma skin cancers—the lighter the skin, the higher the risk[62]. Other types of cancer have also been shown to correlate with geography in Europe[117]. The EuroAIM panels proposed here can therefore help control for stratification in gene-association studies that look for genes functionally involved in increased disease risk, such as different types of cancers. These EuroAIMs here can also be used to control for ancestry in gene-phenotype association studies[39] on the genetic basis of these pigmentation traits; I have shown a few candidate pigmentation genes with high F_{ST} and allele frequency difference among the four samples of European women (Chapter 4).

Table 5.1 – Genetic ancestry tests for Europeans

Company	Product (Markers)	Model (from companies' websites)
DNAPrint	EuroDNA1.0™ (313 autosomal and X SNPs)	1. Northern European subgroup (NOR) 2. Southeastern European (Mediterranean) subgroup (MED) 3. Middle Eastern subgroup (MIDEAS) 4. South Asian subgroup (SA)
DNA Tribes	"BGA Plus" (13 autosomal STRs)	- 19 worldwide regions; European populations subdivided as: 1. Western 2. Eastern 3. Mediterranean 4. Asia Minor - Matches are also given toward the population samples present in the DNA Tribes database
Oxford Ancestors	MatriLine™ (400bp mtDNA sequence)	8 Matrilineal Haplogroups Clans (with overlapping geographic ranges)
	Y-Clan™ (10 Y-STRs)	5 Patrilineal Haplogroups Clans (with overlapping geographic ranges)

Another practical application is recreational genomics, i.e. genetic genealogy. To fill this niche market, a growing number of companies provide genetic tests for various purposes, from disease susceptibility to family relationship, and of course, ancestry[49]. Models of European ancestry proposed by these companies generally agree on a north vs. south pattern but often using different and potentially misleading categories (Table 5.1). For instance, the use of the term "Mediterranean" for a

parental population suggests that all European populations bordering the Mediterranean Sea, from Gibraltar to Greece or possibly Turkey, form an entity which may be quantified as a part of a person's genome. There is, to my knowledge, no scientific evidence of such an entity, which may unfortunately be construed "race" by a layperson browsing the companies' websites—even if they don't use the word "race" explicitly. In the case of the concept of a genetically distinct "Mediterranean" population, the evidence I present with the 10k mapping array clearly shows that there is a significant axis of stratification along the north coast of the Mediterranean Sea, from Greece to Valencia (Spain) through Sicily and the Italian Peninsula. That is, individuals from all these Mediterranean regions alternatively map to any of the three major European poles of ancestry described in Chapter 3 (Iberian, southeastern and northern) and confirmed in Chapter 4. This high differentiation along the northern Mediterranean coast was hinted in previous analyses using 5,000 genome-wide SNPs typed in European individuals[33] and uniparental DNA markers[14]. In Chapter 4, I present an analysis of the largest individual SNP data to date on Europeans of known ancestry, the 317k mapping arrays on 180 individuals. The main population samples represent four corners of Europe, but the results are not as "square" as might be expected from geography. In both PCoA and Bayesian analysis, the principal axis of ancestry distinguishes the two northern population samples from the two southern groups; however the southernmost sample (from Portugal) appears intermediate between the Italian and the Irish samples. The two northern population samples, the Irish and the Polish, even though located in similar latitudes are also separated on the principal clustering axis (PC1 in PCoA). Secondary trends in both PCoA and *structure* confirm an east-west axis in both southern Europe (from Italy to Portugal) and northern Europe (from Poland to Ireland). Therefore the main trend of variation is neither north-south nor east-west, but both at the same time—best seen in Figure 4.3C.

The complexity of the European genetic landscape is not reflected adequately in any model used in commercially available genetic ancestry tests today (Table 5.1), including the EuroDNA 1.0 test presented in Chapter 2. Removing the South Asian and Middle Eastern components from EuroDNA 1.0 (because unlikely parental populations for Europeans) would improve the test; the two remaining components (Northern European and Southern European) would approximately corresponds to the two-cluster model observed and tested in the 10k analysis of Chapter 3 (northern vs. southeastern Europeans). This simple north-southeastern model is the only one I could test, among the three models of European genetic diversity I presented in Chapter 3 (the other two included respectively three and five clusters). However, as further emphasized in Chapter 4, an east-west trend is present in the two top PCs and therefore deserves further investigation. Additionally, the Basque individuals are distinct enough to form their own axis in the 10k array analysis (Chapter 3) and the two

population pools are positioned near a Caucasus population (the Georgian Svans) in the 317k pool analysis (Chapter 4).

In the future, it may be possible to build a European ancestry test along the lines of the model of three or five clusters considered in Chapter 3. Both models are to some extent validated by the 317k population pool analysis, in light of which probably over a dozen-cluster model may also be possible (Chapter 4, Figure 4.2D). The 180-individual analysis of Chapter 4 also showed that more in-depth structure is present and clearly visible with large EuroAIM panels (~ 12k or more), which may be significantly reduced because the 4-cluster pattern among the Irish, Italians, Polish and Portuguese can also be apparent with a few hundred markers (preliminary results, not shown). However it remains to be tested whether such small sets can detect and correct for stratification in Europeans as a whole. LD haplotypes or SNPs in combination with other types of markers such as microsatellites[118] hold promises to reduce the number of markers needed to delineate the landscape of European genetic variation. It is tempting to predict that the same levels of confidence as in the worldwide AncestryByDNA 2.5 test may be reached within a continent relatively as little diverse like Europe.

However, with higher definition of finer details comes greater complexity, and it is questionable whether the concept of parental population can apply to Europe. Indeed no true "parental" population reflecting ancient migration still exists today, and the closest we can find for putative direct descendants of ancestral groups, such as the Basques or Caucasus populations, have been admixed or have drifted. These phenomena will have to be carefully considered before attempting to construct a model for measuring genetic ancestry for the public. The concept of parental population is easily understood in the case of the recent admixture measured by DNAPrint's genetic test AncestryByDNA 2.5, among West Africans, Europeans, East Asians and Indigenous Americans, which is well adapted to the US population. The test becomes harder to interpret in terms of recent admixture when an individual's ancestors do not come from the four parental populations. In the case of Europe, no such model even exists because admixture is more ancient. If the purpose is to model recent population movements, i.e. in the past few centuries, then many more modern European populations need to be considered. If the purpose is to build a model of ancient parental population, the task becomes even more difficult, since populations' histories and ancient admixture at the level of populations' levels need to be researched and accounted for. Such considerations, however, are unlikely to take place in the genetic ancestry business because there is no institution controlling the quality or relevance of genetic tests. One might argue that such a control should not be exercised over a free market; however genetic tests are not neutral products, but rather like movies (which are subject parental guidance) or drugs (which must follow strict standards before coming to market). Genetic tests may have deep

meaning to people's identity as well as health aspects[119]; is it ethically acceptable to risk misleading customers on these issues?

Getting real on human genetic diversity

More generally, an important consideration is the impact of population genomic research on our understanding of human diversity. When scientists debate the apportionment of genetic diversity[50,51,53,55], a range of confusing positions are conveyed to the general public. Researchers working on human genetic variation as well as textbooks justifiably state that biology is not a racist science, adding that the existence of biological races is scientifically discredited, or that racial subdivisions do not reflect genetic discontinuity[53]—although the same scholars may look for patterns of differentiation among populations and point at discontinuities in their research[9,120]. Most anthropology textbooks reflect the view originally advocated by Lewontin[51] that, since more diversity exists within populations than among them, the concept of race is of no genetic significance. As stated more recently in the first publication of the human genome: "Although it may be easy to observe distinct external differences between groups of people, it is more difficult to distinguish such groups genetically, since most genetic variation is found within all groups"[121]. But has this been verified empirically? The fact that some 85% of human genetic variation is accounted for by variation within populations has been confirmed repeatedly[53], but the conclusion that differences among populations are insignificant has been largely contradicted by countless publications on genetic variation at a worldwide level, and for the European populations by the present thesis and recent studies(e.g. [35,46]). This is because the fundamental objection to Lewontin's conclusion is statistical; simply put, "most of the information that distinguishes populations is hidden in the correlation structure of the data and not simply in the variation of the individual factors"[122]. As an illustration, when the question is: "How often is a pair of random individuals from two different populations genetically more similar than a pair of individuals randomly selected from any single population" the classificatory power is less than classification methods using the correlation structure among markers (such as the methods used in this thesis), although the number of loci used is still the most critical variable. Polygenic phenotypes—such as pigmentation traits presented here, are the result of possibly several dozen of markers aggregating within a population and therefore could predict a person's ancestry and vice-versa. In practice it may be argued that such classification decreases predictive power because of environmental and genetics factors[123] and their interplay.

It has also been argued that any observed clustering is an artifact of the discontinuous sampling patterns; that is, if sampling were continuous, only smooth gradients of variation could be observed[124]. However, with proper study design and enough markers, genetic discontinuities can be observed across geographic barriers[4,9] as well as cultural divisions (e.g. in high and low caste Indians, as shown in Chapter 3). Extensive studies of spatial autocorrelation analysis have also shown geographical differences and statistical discontinuities in the distribution of classical genetic markers and physical traits in Europe [9,117,125,126] and other continents[127].

The ~85% of genetic diversity within groups represent an empirical baseline, against which some phenotypes and genomic regions stand out and show high differentiation among populations, in particular those under differential natural selection[42,55]. Consequently, populations do sometimes cluster, and admixture proportions can be measured with AIMs, yielding important practical applications in genealogy, forensics and biomedical research[49,128]—in particular in disease risk assesment[129].

However, interpreting these facts as justification for the concept of "race" (i.e. that genetic differences reflect an individual's worth enforced by social hierarchies) is invalid and dangerous, though not for the reason advanced by Lewontin and others.

Is this racism?

The present attempt at describing and measuring human genetic stratification in Europe has occasionally triggered fear and suspicion in the study subjects, among scientific colleagues and the media. Is it justified to think that genetic ancestry information may cause social oppression and abuse from governments, corporate entities (such as insurance companies), or private agents (such as family members)? How much a difference does it make if we use the word "race" instead of cluster, ethnicity, population, etc..?

It has been shown that *statistical* genetic differences do exist among human groups[123,130], and the work presented here using thousands of markers confirms such discontinuities not only across continental populations, but also within Europe. These genetic discontinuities are typically distributed along (and probably maintained by) boundaries of geography, language or social lines such as the high and low caste in India (Chapter 3). In the past, anthropologists have often presented genetic variation as clines or as phylogenies[11]. The phylogenetic views are generally now avoided because they are better adapted to represent non-interbreeding species. Tree-like representation of human populations, even presented in a rigorous scientific way with proper caveats, can be too easily misinterpreted into

supporting racist ideas of separate hierarchical lineages. Individual representations of human variation which began in the late 1990's[75] are therefore rightfully becoming widespread. The fear that a cluster-like representation could supports racist views is to some extent legitimate; but if clustering does happen, hiding or watering down results will not help eradicating racism.

At the core, what is racism? If it is inherent to our tribally evolved minds to identify kinds to make sense of the world[131]—whether using genetic or cultural traits, can we find ways to reconciliate this function of our brain with humane treatment of all our conspecifics? Unfortunately, the age-old human idea that some groups are superior to others has been accompanied with chronic raiding and warfare throughout human history[132], culminating in devastating world wars and the terrifying applications of eugenics in the 20th century. However, human as we are, we need a representation of the world in kinds (good food v. bad food, dangerous person v. friend, etc.) and if proper identification of the environment meant life or death in the not so distant past, it still has today the non-negligible advantage to help us make sense of our world. The question is not so much whether we should perceive genetic differences as continuous or not (this will depend on the circumstances and study design[4]), but rather, what we will do with that perception. If specific human cultural or physical traits represent danger in our minds, we will act accordingly. Therefore, equally important to the knowledge on individual genetic distances and admixture levels, is what we do with this information.

Even if we can prove that some geographically or linguistically defined groups have a specific set of cognitive or physical abilities superior to that of others, it doesn't justify group eugenics. Egalitarianism is a societal value that should not depend on genetic sameness. For instance, if Neanderthals were still around, they would likely be stronger than us, possibly less gifted on some cognitive levels—and maybe more on others, but nothing would justify abusing them if they were able to hold the same social responsibilities as we do. There is no need to "disprove" the existence of biological subdivisions in the human species to make a case against slavery or any form of institutionalized social oppression of one human group by another—whether these groups are culturally or genetically defined.

Final considerations

The failure of the Human Genome Diversity Project (HGDP) brings up the question of whether efforts at studying human variation should be made at all. Part of the question is about the utility of such an endeavor; the second regards the ethical and moral consideration. Do "racial" drugs such as

BiDil (NitroMed, Lexington, MA) foster the racial views of human kind? Does simply talking about human variation fuel racist policies and behavior?

Our responsibility as scientists studying human variation cannot be taken lightly. Even when avoiding racist terminology, scientific studies can be misinterpreted and abused for racist purposes. We must therefore take a preemptive approach to explain the work we do, and strive to educate the media and public to help them keep pace with new finding and models. More importantly, because the forefront of research is a moving target and full of debates and hypotheses, the degree of certainty of different facets of knowledge must be laid out clearly. We cannot underestimate the naïve perception of the public, and its possible amplification by the media, because the social implication can be heavy; any misconception may cut sources of funding and simply give bad reputation to the entire endeavor of understanding human variation.

Additionally, it has been argued that the harm caused through the use of racial or ethnic categories may outweigh the benefits[72]. Even if a drug such as BiDil has direct health benefits for many African Americans with heart disease, it may cause harm by stigmatizing them. However, eliminating BiDil is not a solution; rather, we must invent a new way to present such products. One promising aspect currently emerging is the use of admixture terminology; not all African Americans have the same amount of West African admixture, yet they all may self-classify categorically. It is an unfortunate fact that, just like in the Jim Crow's days, a person admixed between a European and a West African will be considered African America in this country. In other countries such as Brazil or France, there are specific categories acknowledging various levels of admixture. The emergence of a "mixed" category (in a way, a non-category) is a most positive sign in the celebration of human diversity, which may be the only way to come to terms with the "race" concept.

Note that the words "cluster" and "population" are no substitute for the muddled term "race", which has loaded social and historical aspects. Rather, groupings used in science are proxies for populations meaningful in biomedical research, forensics or anthropology[4,49], and we must carry out the idea that they are just that—temporary and multi-faceted tools, useful only in specific contexts. For instance, the presence of European stratification can create false positives in gene association studies[36,37], but knowing the corresponding genetic affiliation of a European individual will be of no use to predict their social status or ability to play chess. Any human grouping used by science must not be understood in the sense of pure or immutable discrete entity but as the best we can do to represent populations that correlate with genetics in useful ways—just as a palette of discrete colors attempts at describing a rainbow.

REFERENCES

1. Davies N (1998) Europe: A History Harper Perennial
2. Huxley J, Haddon AC (1936) We Europeans : a survey of 'racial' problems. Harper & bros., New York and London
3. Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nature Genet 33:266-275
4. Rosenberg NA, Mahajan S, Ramachandran S, Zhao CF, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. Plos Genetics 1:660-671
5. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381-2385
6. Blumenbach JF (1795) On the Natural Varieties of Mankind: De generis humani varietate nativa. Bergman Publishers, 1969, New York
7. Baum BD (2006) The rise and fall of the Caucasian race: a political history of racial identity University Press, New York
8. Wacholder S, Rothman N, Caporaso N (2002) Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 11:513-520
9. Barbujani G, Sokal RR (1990) Zones of Sharp Genetic Change in Europe Are Also Linguistic Boundaries. Proceedings of the National Academy of Sciences of the United States of America 87:1816-1819
10. Sokal RR, Oden NL, Legendre P, Fortin MJ, Kim JY, Thomson BA, Vaudor A, Harding RM, Barbujani G (1990) Genetics and Language in European Populations. American Naturalist 135:157-175
11. Cavalli-Sforza L, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. Princeton University Press, Princeton NJ
12. Piggott S (1973) Ancient Europe Edinburgh University Press
13. Rightmire GP (1998) Human evolution in the Middle Pleistocene: The role of *Homo heidelbergensis*. Evolutionary Anthropology: Issues, News, and Reviews 6:218-227
14. Jobling M, Hurles M, Tyler-Smith C (2004) Human Evolutionary Genetics. Garland Science
15. Pavlov P, Svendsen JI, Indrelid S (2001) Human presence in the European Arctic nearly 40,000 years ago. Nature 413:64-67
16. Currat M, Excoffier L (2004) Modern humans did not admix with Neanderthals during their range expansion into Europe. PLoS Biol 2:2264-2274
17. Serre D, Langaney A, Chech M, Teschler-Nicola M, Paunovic M, Mennecier P, Hofreiter M, Possnert G, Paabo S (2004) No evidence of neandertal mtDNA contribution to early modern humans. PLoS Biol 2:313-317
18. Caramelli D, Lalueza-Fox C, Vernesi C, Lari M, Casoli A, Mallegni F, Chiarelli B, Dupanloup I, Bertranpetit J, Barbujani G, et al. (2003) Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. PNAS 100:6593-6597
19. Finlayson C, Pacheco FG, Rodriguez-Vidal J, Fa DA, Lopez JMG, Perez AS, Finlayson G, Allue E, Preysler JB, Caceres I, et al. (2006) Late survival of Neanderthals at the southernmost extreme of Europe. Nature 443:850-853
20. Hardy J, Pittman A, Myers A, Gwinn-Hardy K, Fung HC, de Silva R, Hutton M, Duckworth J (2005) Evidence suggesting that Homo neanderthalensis contributed the H2 MAPT haplotype to Homo sapiens. Biochem Soc Trans 33:582-585
21. Evans PD, Mekel-Bobrov N, Vallender EJ, Hudson RR, Lahn BT (2006) Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an

archaic Homo lineage. Proceedings of the National Academy of Sciences of the United States of America 103:18178-18183
22. Cavalli-Sforza L (1998) The DNA revolution in population genetics. Trends in Genetics 14:60
23. Ammerman A, Cavalli-Sforza L (1984) The Neolithic Transition and The Genetics of Populations in Europe. Princeton University Press, Princeton, New Jersey
24. Fort J, Pujol T, Cavalli-Sforza LL (2004) Palaeolithic populations and waves of advance (Human range expansions). Camb Archaeol J 14:53-61
25. Bellwood PS (2005) First Farmers: The Origins of Agricultural Societies. Blackwell Publishing
26. Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426:435-439
27. Currat M, Excoffier L (2005) The effect of the Neolithic expansion on European molecular diversity. Proc R Soc Lond Ser B-Biol Sci 272:679-688
28. Barbujani G, Bertorelle G (2001) Genetics and the population history of Europe. Proc Natl Acad Sci U S A 98:22-25
29. Bellwood P, Renfrew C (2002) Examining the Farming/Language Dispersal Hypothesis. McDonald Institute for Archaeological Research, Cambridge
30. Richards M, Macaulay V, Torroni A, Bandelt HJ (2002) In search of geographical patterns in European mitochondrial DNA. Am J Hum Genet 71:1168-1174
31. Chikhi L, Nichols RA, Barbujani G, Beaumont MA (2002) Y genetic data support the Neolithic demic diffusion model. Proc Natl Acad Sci U S A 99:11008-11013
32. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402-1422
33. Seldin M, Shigeta R, Villoslada P, Selmi C, Klareskog L, Gregersen P (2005) European Population Structure: Ability to distinguish North and South. The American Society of Human Genetics, 55th annual meeting, Salt Lake City, Utah
34. Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, Schurr TG, Zhadanov SI, Osipova LP, Brutsaert TD, et al. (2005) Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Human Genomics 2:81-89
35. Bauchet M, McEvoy B, Pearson LN, Quillen EE , Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD (2007) Measuring European Population Stratification using Microarray Genotype Data. American Journal of Human Genetics 80:948-956
36. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38:904-909
37. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nature Genet 37:868-872
38. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, et al. (2004) Assessing the impact of population stratification on genetic association studies. Nature Genet 36:388-393
39. Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genetic Epidemiology 20:4-16
40. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. American Journal of Human Genetics 72:1492-1504
41. Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika:325–338

42. Shriver MD, Kennedy J, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Human Genomics 1:274-286
43. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567-1587
44. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959
45. Raju NS, Guttman I (1965) A New Working Formula for the Split-Half Reliability Model. Educational and Psychological Measurement 25:963-967
46. Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, Belmont JW, Klareskog L, Gregersen PK (2006) European Population Substructure: Clustering of Northern and Southern Populations. PLoS Genetics 2:e143
47. Parra EJ, Kittles RA, Shriver MD (2004) Implications of correlations between skin color and genetic ancestry for biomedical research. Nature Genet 36:S54-S60
48. Bonilla C, Parra EJ, Pfaff CL, Dios S, Marshall JA, Hamman RF, Ferrell RE, Hoggart CL, McKeigue PM, Shriver MD (2004) Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. Ann Hum Genet 68:139-153
49. Shriver MD, Kittles RA (2004) Genetic ancestry and the search for personalized genetic histories. Nat Rev Genet 5:611-U613
50. Nei M, Roychoud.Ak (1972) Gene Differences between Caucasian, Negro, and Japanese Populations. Science 177:434-&
51. Lewontin RC (1972) The apportionment of human diversity. Evol Biol:381-398
52. Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. Science 201:786-792
53. Barbujani G, Magagni A, Minch E, CavalliSforza LL (1997) An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences of the United States of America 94:4516-4519
54. Rosser Z, al. e (2000) Y-Chromosomal Diversity in Europe Is Clinal and Influenced Primarily by Geography, Rather than by Language. Am J Hum Genet 67:1526–1543
55. Relethford JH (2002) Apportionment of global human genetic diversity based on craniometrics and skin color. American Journal of Physical Anthropology 118:393-398
56. Jablonski NG, Chaplin G (2000) The evolution of human skin coloration. Journal of Human Evolution 39:57-106
57. Miller GF (2000) The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature. Random House, New York
58. Frost P (2006) European hair and eye color - A case of frequency-dependent sexual selection? Evolution and Human Behavior 27:85-103
59. Tasa GL, Murray, C. J. & Boughton, J. M. (1985) Reflectometer reports on human pigmentation. Current Anthropology:511–512
60. Biasutti R (1959) Le razze e i popoli della terra (3rd ed.). Vol. 1. Unione tipografico-editrice torinese, Torino
61. Brace CL, Montagu A (1977) Human Evolution: An Introduction to Biological Anthropology. 2nd ed. Macmillan, New York
62. Lock-Andersen J, Wulf HC, Knudstorp ND (1998) Interdependence of eye and hair colour, skin type and skin pigmentation in a Caucasian population. Acta Dermato-Venereologica 78:214-219
63. Makova K, Norton H (2005) Worldwide polymorphism at the MC1R locus and normal pigmentation variation in humans. Peptides 26:1901-1908

64. Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C, Sajantila A, Jackson IJ, Birch-Machin MA, et al. (2000) Evidence for variable selective pressures at MC1R. American Journal of Human Genetics 66:1351-1361
65. Bonilla C, Shriver MD, Parra EJ, Jones A, Fernandez JR (2004) Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. Human Genetics 115:57-68
66. Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, et al. (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112:387-399
67. Rees JL (2003) Genetics of hair and skin color. Annual Review of Genetics 37:67-90
68. Sturm RA, Frudakis TN (2004) Eye colour: portals into pigmentation genes and ancestry. Trends in Genetics 20:327-332
69. Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK (2003) A classifier for the SNP-Based inference of ancestry. Journal of Forensic Sciences 48:771-782
70. Beals RL, & Hoijer, H. (1965) An introduction to anthropology (3rd ed.). Macmillan, New York

71. Kalinowski ST, Wagner AP, Taper ML (2006) ML-RELATE: a computer program for maximum likelihood estimation of relatedness and relationship. Molecular Ecology Notes 6:576-579
72. Ossorio PN (2006) About face: Forensic genetic testing for race and visible traits. Journal of Law Medicine & Ethics 34:277-+
73. Halder I, Shriver M, Thomas M, Fernandez J, Frudakis T (in prep.) A panel of Ancestry Informative Markers for estimating individual BioGeographical ancestry and admixture from four continents: utility and applications.
74. Chakraborty R, Jin L (1993) A unified approach to study hypervariable polymorphisms: Statistical considerations of determining relatedness and population distances. In: Pena SDJ, Chakraborty R, Epplen J, Jeffreys AJ (eds) DNA Fingerprinting: Current State of the Science. Birkhauser, Basel, pp 153-175
75. Mountain JL, Cavalli-Sforza LL (1997) Multilocus genotypes, a tree of individuals, and human evolutionary history. Am J Hum Genet 61:705-718
76. Wright S (1951) The genetical structure of populations. Annals of Eugenics 15:323-354
77. Akey JM, Zhang G, Zhang K, Jin L, Shriver M (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805-1814
78. Chae SS, Warde WD (2006) Effect of using principal coordinates and principal components on retrieval of clusters. Computational Statistics & Data Analysis 50:1407-1417
79. Chessel D, Dufour A-B, Thioulouse J (2004) The ade4 package-I- One-table methods. R News:5-10.
80. Golden CJ, Fross KH, Graber B (1981) Split-Half Reliability and Item-Scale Consistency of the Luria-Nebraska Neuropsychological Battery. Journal of Consulting and Clinical Psychology 49:304-305
81. Lord FM (1956) Sampling Error Due to Choice of Split in Split-Half Reliability Coefficients. Journal of Experimental Education 24:245-249
82. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS (2002) Truncated product method for combining P-values. Genetic Epidemiology 22:170-185
83. Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Molecular Ecology Notes 4:137-138
84. R_Development_Core_Team (2006) R: A Language and Environment for Statistical Computing, Vienna, Austria

85. Bauchet M, McEvoy B, Pearson LN. , Quillen EE. , Tamara Sarkisian, Kristine Hovhannesyan, Ranjan Deka, Daniel G. Bradley, Shriver MD (in press) Measuring European Population Stratification using Microarray Genotype Data. American Journal of Human Genetics
86. Kennedy GC, Matsuzaki H, Dong SL, Liu WM, Huang J, Liu GY, Xu X, Cao MQ, Chen WW, Zhang J, et al. (2003) Large-scale genotyping of complex DNA. Nat Biotechnol 21:1233-1237
87. Saitou N, Nei M (1987) The Neighbor-Joining Method - A New Method For Reconstructing Phylogenetic Trees. Mol Biol Evol 4:406-425
88. Mountain JL, CavalliSforza LL (1997) Multilocus genotypes, a tree of individuals, and human evolutionary history. Am J Hum Genet 61:705-718
89. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997-1004
90. Belle EMS, Landry PA, Barbujani G (2006) Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. Proceedings of the Royal Society B-Biological Sciences 273:1595-1602
91. Pereira L, Richards M, Goios A, Alonso A, Albarran C, Garcia O, Behar DM, Golge M, Hatina J, Al-Gazali L, et al. (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Research 15:19-24
92. Mantel N (1967) Detection of Disease Clustering and a Generalized Regression Approach. Cancer Research 27:209-220
93. McEvoy B, Richards M, Forster P, Bradley DG (2004) The Longue duree of genetic ancestry: Multiple genetic marker systems and Celtic origins on the Atlantic facade of Europe. Am J Hum Genet 75:693-702
94. Bauduer F, Feingold J, Lacombe D (2005) The Basques: Review of population genetics and Mendelian disorders. Human Biology 77:619-637
95. Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. American Journal of Human Genetics 62:1171-1179
96. Clackson J (1994) The Linguistic Relationship Between Armenian and Greek. Vol. No 30. Publications of the Philological Society, London
97. Weaver TD, Roseman CC (2005) Ancient DNA, late neandertal survival, and modern-human-Neandertal genetic admixture. Curr Anthropol 46:677-683
98. Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K (2005) An Icelandic example of the impact of population structure on association studies. Nature Genetics 37:90-95
99. Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. Ann Hum Genet 66:393-405
100. Gunderson KL, Steemers FJ, Ren HG, Ng P, Zhou LX, Tsan C, Chang WH, Bullis D, Musmacker J, King C, et al. (2006) Whole-genome genotyping. DNA Microarrays Part a: Array Platforms and Wet-Bench Protocols, pp 359-+
101. Weir BS, Cockerham CC (1984) Estimating F-Statistics For The Analysis Of Population-Structure. Evolution 38:1358-1370
102. Pechenizkiy M, Puuronen S, Tsymbal A (2006) On Combining Principal Components with Parametric LDA-based Feature Extraction for Supervised Learning. Foundations of Computing and Decision Sciences 31:59-73
103. McEvoy B, Beleza S, Shriver MD (2006) The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. Human Molecular Genetics 15:R176-R181
104. Sanchez-Velasco P, Leyva-Cobian F (2001) The HLA class I and class II allele frequencies studied at the DNA level in the Svanetian population (Upper Caucasus) and their relationships to Western European populations. Tissue Antigens 58:223-233

105. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, et al. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75:910-918
106. Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A (2002) The making of the African mtDNA landscape. American Journal of Human Genetics 71:1082-1111
107. Lamason RL, Mohideen M, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao XY, Humphreville VR, Humbert JE, et al. (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310:1782-1786
108. Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA (2007) A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. American Journal of Human Genetics 80:241-252
109. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. American Journal of Human Genetics 74:1111-1120
110. Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, Bertorelle G, Chamberlain AT, Nunes TP, Metodiev S, Ferrand N, et al. (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. Nature Genet 35:311-313
111. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics 39:31-40
112. Box GEP (1979) Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN (eds) Robustness in Statistics. Academic Press, New York
113. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, et al. (2004) Methods for high-density admixture mapping of disease genes. American Journal of Human Genetics 74:979-1000
114. Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, Kidd KK (2005) Linkage disequilibrium patterns vary substantially among populations. European Journal of Human Genetics 13:677-686
115. Chakraborty R, Weiss KM (1988) Admixture As A Tool For Finding Linked Genes And Detecting That Difference From Allelic Association Between Loci. Proc Natl Acad Sci U S A 85:9119-9123
116. Shriver M, Frudakis T, Budowle B (2005) Getting the science and the ethics right in forensic genetics. Nature Genet 37:449-450
117. Rosenberg MS, Sokal RR, Oden NL, DiGiovanni D (1999) Spatial autocorrelation of cancer in western Europe. European Journal of Epidemiology 15:15-22
118. Mountain JL, Knight A, Jobin M, Gignoux C, Miller A, Lin AA, Underhill PA (2002) SNPSTRs: Empirically derived, rapidly typed, autosomal Haplotypes for inference of population history and mutational processes. Genome Research 12:1766-1772
119. Marks J (2005) New Information, Enduring Questions. GeneWatch 18:11-16
120. Barbujani G, Oden NL, Sokal RR (1989) Detecting Regions of Abrupt Change in Maps of Biological Variables. Systematic Zoology 38:376-389
121. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921
122. Edwards AWF (2003) Human genetic diversity: Lewontin's fallacy. Bioessays 25:798-801

123. Witherspoon D.J. , Wooding S. , Rogers A.R., Marchani E.E. , Watkins W.S. , Batzer M.A. , L.B. J (2007) Genetic Similarities Within and Between Human Populations. Genetics [Epub ahead of print]
124. Serre D, Paabo SP (2004) Evidence for gradients of human genetic diversity within and among continents. Genome Research 14:1679-1685
125. Falsetti AB, Sokal RR (1993) Genetic-Structure of Human-Populations in the British-Isles. Annals of Human Biology 20:215-229
126. Sokal RR, Uytterschaut H (1987) Cranial Variation in European Populations - a Spatial Autocorrelation Study at 3 Time Periods. American Journal of Physical Anthropology 74:21-38
127. Sokal RR, Thomson BA (1998) Spatial genetic structure of human populations in Japan. Human Biology 70:1-22
128. Berg K, Bonham V, Boyer J, Brody L, Brooks L, Collins F, Guttmacher A, McEwen J, Muenke M, Olson S, et al. (2005) The use of racial, ethnic, and ancestral categories in human genetics research. American Journal of Human Genetics 77:519-532
129. Kittles RA, Weiss KM (2003) Race, ancestry, and genes: Implications for defining disease risk. Annual Review of Genomics and Human Genetics 4:33-67
130. Long JC, Kittles RA (2003) Human genetic diversity and the nonexistence of biological races. Human Biology 75:449-471
131. Berreby D (2005) Us and Them: Understanding Your Tribal Mind Little, Brown and Company
132. Keeley LH (1996) War before Civilization: The myth of the peaceful savage. Oxford University Press, New York

APPENDIX A

Consent Form

LabID#: _____

Password: _____

Name (print please): _____

CONSENT FORM
CONSENT TO ACT AS A SUBJECT IN AN EXPERIMENTAL STUDY

TITLE: **Genetics of Human Pigmentation, Ancestry and Facial Features**

Principal Investigator: Mark D. Shriver, Ph.D.
Department of Anthropology
Penn State University
512 Carpenter Bldg.
University Park, PA 16802
814-863-1078

Co-Investigators:

Marc Bauchet, Xianyun Mao, Abby Bigham,
Ellen Quillen, Laurel Pearson, Denise Liberton
Department of Anthropology
Penn State University
409 Carpenter Bldg.
University Park, PA 16802
814-865-2313

Tony Frudakis, Ph.D.
Chief Scientific Officer
DNAPrint Genomics, Inc.
900 Cocoanut Ave.
Sarasota, FL 34236
941-366-3400

Samuel Richards, Ph.D.
Department of Sociology
Penn State University
0211 Oswald Tower
University Park, PA 16802
814-863-7456

Description: Dr. Shriver and colleagues at Penn State University and DNAPrint Genomics, Inc. are collecting information to use to identify the genes that determine normal variation in a number of common traits. These traits include genetic ancestry, hair, eye, and skin color, and facial features. In addition to studying the genes for common physical traits, these researchers will also study genetic variation across the genome and the processes of admixture, the mixing of populations that were previously separated. These researchers plan to collect samples from a total of 8,000 persons from throughout the world.

Just so that you are aware, Dr. Shriver is a paid scientific consultant for DNAPrint Genomics, Inc., the company that performs the Ancestry 2.5 test. Given he is both a consultant for this commercial company and a Professor at Penn State University, he is monitored by the PSU Office of Research Protections to prevent unacceptable conflicts of interest.

If you have been invited to participate in this study through a class at Penn State, note that your participation or lack thereof will not affect your grade in any way. You are free to participate or not as you chose.

The total time of your participation is approximately 90 minutes. You are being asked to be a volunteer in this study. If you decide to participate, here is what will happen:

You will be asked to complete a form with your name, and place and date of birth of yourself, your parents and about your family's ethnic and racial ancestry. You will then be asked to use two cheek brushes to gently collect cells from the insides of your cheeks and to undergo fingerstick blood collection, where a sharp lancet will be used to prick your index or middle finger and the resulting blood drops collected on a special DNA paper card. Your finger will be cleaned with an alcohol wipe before using the lancet and a Band-Aid will be applied afterwards. You will then have five standard digital photographs taken; one of your face smiling, one neutral, two profiles and one close up of your eyes and two 3d photographs: one smiling and one neutral. You will also have your skin color and hair color measured using a handheld reflectometer. The reflectometer shines two colored lights on your skin and measures the amount of light reflected back.

The facial photographs (both standard 2d and 3d photos) have three uses: 1) research on the genetics of facial traits and the ability of observers to detect ancestry contributions, 2) inclusion in a DNAPrint forensic database from where they may be made available to police as examples of individuals who have particular levels of genetic ancestry and on research regarding the utility of such tests and 3) educational purposes to demonstrate the appearance of persons with particular ancestry levels. All participants in this study consent to the first two uses of the photographs. Once you have had the opportunity to review your Ancestry 2.5 results, you will be given the opportunity to provide additional consent for the use of your photo and ancestry results together in public. There is a separate consent form for this and it will be up to you as to whether you wish to consent or not. The photographs will be stored in a secure password protected computers in locked offices. Given the relative novelty of the 3D photographs you are given the opportunity to receive a copy of your 3D photos on a CD along with the software for viewing them on Windows computers. Please check YES if you wish to receive such a CD, in which case it will be available for you to pick up in 508 Carpenter no later than two days after this session, and until the end of the semester.

___ YES, please prepare a 3D photo CD for me.
___ NO, I'd rather not have my 3D photos given to me.

Your Ancestry 2.5 results will be returned to you via email in an encrypted file. Your randomly generated password is indicated on the top of this consent form, so please keep this form. If you lose it, you can contact Shriver and ask for another copy. Once you have had a chance to review your results you will be invited to attend a second optional meeting of approximately 90 mins to ask questions regarding the interpretation of your results. At this time you will be invited to consent to a broader use of you photo and ancestry results. These uses are for presentation in scientific and public forums such as books, manuscripts, posters, or slide presentations. For all three of these uses your name will not be connected to your photo.

Currently, a definite study completion date is unknown but we expect to continue this work for a number of years. Your samples will not be made available for any other research and at the completion of this study, or the year 2035 at the latest, samples of your blood and DNA will be destroyed. Copies of your photographs that are contained in the forensic database and those that have been published, will not be destroyed, but all other data and materials will be.

Risks and Benefits: You will be provided with sterile cheek brushes with a soft foam pad. The chance on injury using a cheek brush is minimal. There is a chance of soreness and in rare cases infection at the site of the fingerstick blood draw. To minimize the risk of infection we will take standard safety precautions in drawing fingerstick blood. Specifically, participants will be seated, have their fingers cleaned with a fresh alcohol pad, pricked with a sterile lancet and have their blood drops collected on a fresh DNA paper card. The potential benefits of your participation are that you may help scientists to better understand the biology that underlies these important and interesting physical features and that by enhancing the DNAPrint forensic database, there is the possibility that your contribution may assist in solving important crimes. Although your photograph may be made available to criminal investigators along with your ancestry profiles, these are investigational tools that will only be used internally by police. Although the police may be able to view your photograph, there will not be casual access to your name, which will not be sent to DNAPrint. However, in the unlikely case that you were recognized as a suspect, you should be aware that a court ordered subpoena could be issued in which case your name and other identifying information would be provided to police. Additionally, you will be provided with your DNAPrint Ancestry 2.5 test results that may provide interesting clues regarding your genetic heritage. You should be aware that the ancestry results may be surprising and might possibly provide indications of unexpected ancestral contributions. For example, persons expecting only ancestry from one continental population may get a result showing more than one contribution. Likewise some who have a familial oral tradition or other evidence of particular ancestral contributions may have results that contradict these expectations. Although the DNAPrint test is firmly based on scientific evidence, the results are presented in a probabilistic framework and should be interpreted accordingly. Ideally, any genetic ancestry information is taken in the context of broader information such as genealogical research and oral records.

Costs and Payments: You understand that you will not be charged for any of the testing in this study. The DNAPrint Ancestry 2.5 test retails for $219, but will be provided to you at no cost.

Confidentiality: The records and data files related to this research project will be maintained in the Department of Anthropology and only personnel directly associated with this project will have access to them. In order to monitor this research project, the following may review and copy records related to this research; The Office of Human Research Protections in the U.S. Department of Health and Human Services; The Penn State University Biomedical Institutional Review Board; The Penn State University Office for Research Protections of Penn State. You consent to publication of any information for scientific purposes as long as your identity will not be revealed outside of the uses of

the photographs as indicated above. The genetic testing which will be done is strictly for research purposes we will not be testing for genes that may predispose to disease.

Right to Withdraw: You do not have to take part in this study and can withdraw at any time if you change your mind. If you do withdraw, all samples relevant to you will be destroyed. You also have the right to refuse to answer any particular questions. You also understand that you may be removed from the study by the investigators in the event that it has been determined that the information you have provided is incorrect.

Compensation for Illness or Injury: In the unlikely event that you should be injured as a result of your participation in this study, you understand that neither financial compensation nor free medical treatment is provided. You also understand that you are not waiving any rights that you may have against the University for injury resulting from negligence of the University or investigators.

Future contact: We will be performing other research with volunteers from this study. Can we contact you about participating? _____YES _____ NO

Voluntary Consent: Any questions I have about this research will be answered by Dr. Shriver whom I may call at (814) 863-1078 or contact by email at mds17@psu.edu. If I have questions about my rights as a research participant, I can contact the PSU Office for Research Protections at (814) 865-1775. By signing this form, I agree that I am at least 18 years of age and to participate in this study.

_____	_____
Date and Time	Participant's Signature
_____	_____
Date and Time	Person Obtaining Consent

APPENDIX B

Coriell Samples Detail

http://www.anthro.psu.edu/biolab/AppendixB--Coriell.Details.xls

APPENDIX C

DNAPrint EuroDNA 1.0 AIMs

http://www.anthro.psu.edu/biolab/AppendixC--DNAPrintAIMs.xls

APPENDIX D

Example EuroDNA 1.0 Test Result

http://www.anthro.psu.edu/biolab/AppendixD--BNC50143MS-EURO.pdf
(password = 1111)

APPENDIX E

180 Individuals Typed on 317k Arrays: Ancestry and Phenotypes

http://www.anthro.psu.edu/biolab/AppendixE--4x45phenoptypes.xls

VITA

Marc P. Bauchet

EDUCATION

- 2003-2007 Ph.D. in Biological Anthropology at The Pennsylvania State University.
- 1998-2001 Anthropology and Archaeology Master student at Harvard Extension School
- 1989-1994 Masters in Engineering (electronics, data and image processing) at ICPI (Institut de Chimie et Physique Industrielles) Lyon, France

WORK EXPERIENCE

- 2003-2007 The Pennsylvania State University:
  - Research Assistant in the Shriver Laboratory (Anthropological Genomics)
  - Teaching Assistant (Introduction to Physical Anthropology)

- 1994-2003 Professional software developer, engineer and project manager:
  - 2001-2003 A2iA (New York, NY)
  - 1997-2001 Dragon Systems (Boston, MA)
  - 1994-1997 Thomson Multimedia (France)
  - 1994 Thomson Brandt (Germany)

FIELDWORK

- 2004-2006 Organized and participated in anthropometric data collection at University Park PA, in France, Italy, Poland, Portugal and Ireland, for Shriver Lab project: "Genetics of Human Pigmentation, Ancestry and Facial Features"
- Summer 2001 Earthwatch Institute excavation with Dr. Michael Walker in Murcia, Spain

AWARDS AND MEMBERSHIPS

- 2007 Baker Fund Award
- 2006-2007 Weiss Fellowship: tuition and salary
- 2006 Hill Fellowship Award
- 2003, 2004 Weiss Fellowship: tuition and salary

American Society of Human Genetics, member

SELECTED PRESENTATIONS AND PUBLICATIONS

- Bauchet M, McEvoy B, Pearson LN, Quillen EE , Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD (2007) Measuring European Population Stratification using Microarray Genotype Data. *American Journal of Human Genetics* 80:948-956
- "Human Genomic Landscapes in Europe" Invited seminar presentation at the Reich Lab, Harvard Medical School (June 2006) and Max Planck institute, Germany (January 2007)
- "The use of biogeographical ancestry for forensic, biomedical, and recreational genomics" Paper and presentation at 75th annual American Association of Physical Anthropologists meeting, 8-11 March 2006