CHARACTERIZATION OF GENETIC VARIATION IN NORTH AFRICAN AND SPANISH POPULATIONS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Wei Wang

January 2011

© 2011 Wei Wang

ABSTRACT

High-density Single Nucleotide Polymorphism (SNP) scans of the human genome have been applied in many populations worldwide to investigate their genetic characteristics. However, populations in North Africa, an isolated subcontinental area between Sub-Sahara Africa and Europe, have not been examined. In the present study, seven North African populations and four neighboring Spanish populations are analyzed using a high-density SNP microarray. North African populations appear to form a clinal pattern of genetic differentiation between Sub-Saharan Africans and Europeans, being much more similar to Europeans than to Sub-Saharans. The genetic similarity between North African populations exhibits an east-west gradient pattern corresponding to their geographic locations. High and varying levels of autozygosity, as well as a potentially indigenous genetic component, are observed in North Africans. Noticeably, Tunisians turn out to be the North African population most distinct from Europeans and Sub-Saharans, and have the highest levels of autozygosity. Basques can be clearly distinguished from other Spanish populations, as being more similar to the Western Europeans, and also have the largest number of fixed ancestral and derived alleles among all the populations studied. The ancestral allele frequency distribution of Basques is most similar to that of East Asians, suggesting a small effective population size. All these results indicate that the Basque population is a genetic isolate distinct from the surrounding Spanish populations as well as from other Southern and Western European populations, although the magnitude of genetic differentiation is subtle.

BIOGRAPHICAL SKETCH

Wei Wang was born in Harbin, China. He got a Bachelor of Science in biology from Zhejiang University, Hangzhou, in 1991, and a Master of Science in genetics in Harbin Medical University, Harbin, in 1994. Then he pursued a Ph.D. in molecular and cell biology at the State University of New York–Downstate Medical Center, in Brooklyn, New York, and earned the degree in 1999. After that he worked at New York University School of Medicine as a research assistant professor in the Department of Obstetrics and Gynecology, to develop SNP genotyping methods for a genetic epidemiology study for about two years. In 2001, he started to work as a bioinformatics scientist, using DNA microarray to discover cancer biomarkers and develop diagnostics tests for Arcturus Applied Genomics in San Diego, California, and later for Arcturus Bioscience in Mountain View, California. In 2005, he came to Cornell University as the director of the Microarray Core Facility to supervise the facility and provide consultation and data analysis on microarray to the life science researchers. He then participated in the Cornell Employee Degree Program to pursue a Master of Science in Biometry with Dr. Carlos Bustamante in the Department of Biological Statistics and Computational Biology, with a focus on the genome-wide genetic variation study of populations.

[Dedicated to my whole family]

ACKNOWLEDGMENTS

I would like to thank many people upon the completion of my M.S. study in biometry. First and foremost, my thanks go to my family. Pursuing a graduate degree while working full time is a tough task, requiring most of my spare time. The support from my family has been unconditional and endless; I could not have accomplished this without them. My sincerest gratitude goes to my thesis adviser, Dr. Carlos Bustamante, for his guidance and support, and his insight and knowledge throughout my thesis research, as well as for the flexibility he gave me to work at my own pace. I would like to offer my gratitude to all members of the Bustamante research group for their help with every aspect of the work, in addition to their friendship. I feel fortunate to have had the opportunity to be part of this first-class research group on population genetics. I would also like to thank other members of my thesis committee, Drs. Andrew Clark and Jason Mezey, for their advice, support, and suggestion on my thesis. Finally, I would like to acknowledge the collaborators in Spain, Drs. David Comas, Jaume Bertranpetit, and Laura Rodriguez of the Institute of Evolutionary Biology at Barcelona, for providing all the human DNA samples used in the present study, as well as suggestions on the analysis strategy of the research project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAF:	Ancestral allele frequency
AMH:	Anatomically Modern Human
GWAS:	Genome-Wide Association Study
HWE:	Hardy-Weinburg Equilibrium
HGDP:	Human Genome Diversity Project
IBS:	Identical by State
LD:	Linkage Disequilibrium
MAF:	Minor Allele Frequency
MDS:	Multidimentional Scaling
mtDNA:	Mitochondrial DNA
NRY:	Non-recombining Region of Y chromosome
PCA:	Principal Component Analysis
QC:	Quality Control
ROH:	Runs of extended Homozygosity
SNP:	Single Nucleotide Polymorphism

CHAPTER 1

INTRODUCTION

Recent advancement in DNA genotyping and sequencing techniques has revolutionized human population genetics research. High-density SNP genotyping microarrays provide a powerful and affordable tool to investigate the genome-wide pattern of genetic variation across a large number of individuals. Such large-scale surveys of genomic variation at high resolution dramatically expand our understanding of aspects of human evolutionary history, such as migration, change in effective population size, range expansion, and adaptation (International HapMap Consortium, 2007; Sabeti et al., 2007; Nielsen et al., 2007; Li et al., 2008; Jakobsson et al., 2008; HUGO Pan-Asian SNP Consortium, 2009). Large-scale genotyping projects have also been carried out to elucidate the genetic basis of complex diseases using the Genome-Wide Association Study (GWAS) strategy. Up until July 30, 2010, 608 publications about the Genome-Wide Association Study have been cataloged at the National Institutes of Health's National Human Genome Research Institute, (Hindorff et al., 2009; available at www.genome.gov/gwastudies).

DNA sequence variation in human genomes has been extensively characterized in a few populations representing major continental groups in the International HapMap Project (International HapMap Consortium, 2005 and 2007). Great effort has been made to discover SNPs by DNA resequencing and to determine SNP allele frequency by microarray genotyping in populations of Western European, Sub-Saharan African, and East Asian ancestry. Based on these findings, large number of worldwide populations have been characterized by microarray SNP genotyping in the

Human Genome Diversity Project (Li et al., 2008; Jakobsson et al., 2008). There are also many other genetic diversity studies focusing on the fine-scale population substructure within continents, such as in Europe (Novembre et al., 2008; Tian et al., 2008; Nelis et al., 2009), Asia (HUGO Pan-Asian SNP Consortium 2009; Teo et al., 2009), and Africa (Tishkoff et al., 2009). However, the genetic variations and relationships among populations of North Africa, and the neighboring Iberian populations across the Strait of Gibraltar, have not been well characterized on a genome-wide level. This geographical region is at the junction between African and Eurasian continents and harbors rich culture and ethnic diversity. The present study aims to fill in this gap of knowledge by genotyping seven North African populations and four Spanish populations with high-density SNP microarray, to explore the genetic relationship among populations within this region and with other major worldwide populations.

1.1 North Africa and the Evolution of Modern Humans

North Africa includes seven countries or territories in the north-most region of the African continent: Egypt, Libya, Tunisia, Algeria, Morocco, Western Sahara, and Sudan. The North African region is effectively isolated from the rest of the African continent by the Sahara Desert in the south, and separated from the European continent by the Mediterranean Sea in the north. This anthropological island, isolated by huge geographical barriers, is only narrowly connected to the Middle East by the Sinai Peninsula of Egypt, and possibly also to the Iberian Peninsula through the Strait of Gibraltar in prehistoric times. Although North Africa is ecologically and culturally isolated from the rest of Africa, this region has been historically influenced by seafaring civilizations, such as Greeks and Romans, which could travel across the

Mediterranean Sea, and later was invaded and conquered by the Islamic Arabs from Southwestern Asia in the seventh century. Therefore, North Africa had more culture exchange, and migration, with the Middle East and Europe than with Sub-Saharan Africa. Berbers are believed to be the indigenous inhabitants of the western part of North Africa, also called Maghreb, while the eastern part of North Africa has been inhabited by Egyptians, mostly along the Nile Valley. Arabic and Berber, both of which belong to the Afro-Asiatic language family, are the languages commonly used in the North Africa region.

Archaeological evidence indicates that modern humans were present in North Africa as early as 45,000 years ago in the form of Aterian industry, although the continuity of human occupancy in this region still needs more support (Garcea and Giraudi, 2006). The earliest fossil evidence of the anatomically modern human (AMH) phenotype has been discovered in Ethiopia in East Africa, which is dated back approximately 130,000 to 195,000 years ago (Day, 1969). It is generally accepted, as supported by fossil and archaeological evidence, that modern humans originated from a small isolated population in Africa up to 2 million years ago, during the Late Pleistocene stage (Walter et al., 2000; Clark et al., 2003; White et al., 2003). This ancestral population is thought to have undergone dramatic growth and range expansion throughout the Old World, and it then completely replaced the archaic forms of other Hominin lineages (Cann et al., 1987; Harpending et al., 1998; Excoffier et al., 2002).

In the next epoch of evolution, long after their origination in Africa and their assimilation of other archaic lineages, anatomically modern humans underwent a global diaspora. An "Out-of-Africa" model is gaining wide acceptance regarding the recent dispersal of modern humans throughout the whole world around 50,000 to 60,000 years ago (Foster and Matsumura, 2005; Mellars, 2006; Torroni et al., 2006).

Two possible routes have been proposed for this long-range migration between continents. The "southern route" of migration is proposed to start from the Horn of Africa along the coast of the Persian/Arabian Gulf, and continue farther to the Indian Ocean, reaching Southeast Asia and Australia, then radiating farther to East Asia and finally to the Americas. It is gaining more evidence for support and favored over the "northern route" of dispersal, which is proposed to be a land route going eastward through the Levant and across the Eurasian Steppe, then turning south through the Asian mainland (Mellars, 2006; Olivieri et al., 2006; Torroni et al., 2006; HUGO Pan-Asian SNP Consortium, 2009).

According to the Out-of-Africa model of recent migration, the delayed settlement in most parts of West Eurasia is thought to have resulted from an offshoot of the east-bound coastal migration route (Macauley et al., 2005; Mellars, 2006). Paleoenvironmental evidence supports the likelihood that the ancestors of West Eurasians experienced a lengthy pause in migration after their initial settlement, probably in the Middle East region along the Persian/Arabian Gulf, until climate improvement allowed them to further expand northward to the Levant and then Europe (Van Andel et al., 1996). Such a back-immigration from West Asia is not likely to have taken place until about 50,000 years ago, when the wetter climate reduced the size of the expansive desert extending from North Africa to Central Asia.

Today North Africa is isolated from the rest of Africa by the Sahara Desert, but connected to the West Asia near its junction with East Africa. It is generally thought that colonization of modern humans in North Africa is parallel to that in Europe, most likely from West Asia during the same period of time, around the early Upper Paleolithic. Therefore, characterization of genetic variation in contemporary North African populations and comparison to populations in other geographic regions

can help one obtain a better understanding of the pattern and timing of the latest out-of-Africa migration, especially the peopling of West Eurasian regions.

1.2 Population Genetics of North Africans

Until recently, population genetics studies of North Africans, as well as other ethnic groups, have mostly been relying on the uniparental haploid DNA sequence variations, namely mitochondrial DNA (mtDNA) and the non-recombining region of Y chromosome (NRY). The haploid genetic markers can be transmitted from only one of the parents and are not subjected to recombination, therefore new mutations are accumulated in a sequential manner in the radiating lineages. Thus, the resulting sequence divergence on mtDNA or Y chromosome over the course of time gives rise to monophyletic haplogroups, which can be used to reconstruct the phylogeny of all lineages in a straightforward way (Wallace 1995). This type of DNA sequence differentiation happens during the process of human migration into different geographical regions, so the sequentially generated haplogroups and subhaplogroups tend to be enriched or even limited to certain populations in specific geographic areas. Thus, these uniparental genetic markers provide a simple and reliable system to trace back the human migration pattern around the world. As the mtDNA has very high sequence evolution rate, at least one order of magnitude higher than nuclear chromosomal DNA (Neckelmann et al., 1987; Wallace et al., 1987), the haplogroups of mtDNA are more diverse and informative than those of the non-recombining region of Y chromosome.

The worldwide phylogenic tree of human mtDNA demonstrates that the root layer L split into a series of branches, L0, L1, through L5. All of these haplogroups are African-specific sequences, except a more peripheral haplogroup, L3, which is

shared between Africans and populations in the rest of the world (Underhill and Kivisild, 2007; Olivieri et al., 2006). The L3 haplogroup further splits into a number of subclades that are exclusively present in African populations, as well as two other branches, M and N, which give rise to all non-African mtDNA lineages around the world. The R haplogroup, an early derivative of N, is also considered an extant founder haplogroup of non-Africans as it is widely dispersed in non-African populations and splits into many subclades of high-sequence variation. In summary, the first informative split of mtDNA phylogenic tree is at the level of L3 and M/N/R clades to distinguish non-African from African populations. The subsequent split in mtDNA tree beneath M, N, and R haplogroups is informative to distinguish all major continental regions except the Americas.

European and Near Eastern populations primarily carry the maternally inherited mtDNA from N-derived haplogroups N1, N2a, W, X, and R-derived haplogroups R0 (including R0a, H, and V), J, T, and U (except U6). Haplogroup U is nested in haplogroup R; it is broadly distributed in wide range of geographic regions, from North Africa and Europe to Central and South Asia, at a very high overall frequency of 15%–30% (Richards et al., 2000; Kivisild et al., 2003; Quintana-Murci et al., 2004). Based on the complete mtDNA, the most informative mtDNA haplogroup for North Africa is U6, which is almost exclusively found in the North African populations. It has been proposed that the U6 lineages represent a return of the West Asian branch to North Africa after the out-of-Africa exit, possibly around 39,000 to 52,000 years ago (Maca-Meyer et al., 2001). Detailed analysis of the U6 lineage by complete sequencing suggests that the Near East is the most likely origin of the proto-U6 haplogroup, which spread to North Africa around 30,000 years ago (Maca-Meyer et al., 2003). Successive expansion is revealed by various subhaplogroups of U6. It has been proposed that subclade U6a signals the range expansion from Northwestern

Africa to East Africa in Paleolithic times, while its derivative U6a1 reflects the posterior return to Maghreb. The U6b and U6c subhaplogroups represent more localized expansion in West Africa, possibly reaching as far as Iberian Peninsula and Canary Islands in prehistoric times. One interesting finding is that Berbers of North Africa and Sami, hunter-gatherers of Scandinavia, share an extremely young subhaplogroup U5b branch that is only approximately 9,000 years old (Alessandro et al., 2005). This reveals a direct maternal connection between the two contemporary populations far away from one another, confirming that the southwestern refuge area of Europe is the source of late-glacial expansion to repopulate northern Europe after the Last Glacial Maximum.

Besides U6, haplogroups X1 and M1 have also been reported to be informative for North Africa populations. Haplogroup X can be further divided into two major subclades X1 and X2 based on complete mtDNA sequence. The distribution of subhaplogroup X1 is restricted to North Africa, East Africa, and the Near East, while the diversity of X1 indicates these lineages coalescenced at an early time, most likely in North Africa (Reidla et al., 2003). Subhaplogroup X2 obviously underwent a recent population expansion in Eurasia, probably around or after the Last Glacial Maximum, and is currently distributed in a wide range at Europe, the Near East, Western and Central Asia, and North Africa.

The M1 haplogroup is the only daughter clade of super-haplogroup M detected in Africa, and seems to be predominantly specific to Africa. The M1 haplogroup exhibits high frequencies in East Africa, but is also observed in North Africa (Olivieri et al., 2006). This unique geographical distribution of the M1 haplogroup brings up the question of whether it originated before or after the initial out-of-Africa expansion. Phylogeographic studies of the entire and partial mtDNA sequence point out that the coalescent time of the African M1 haplogroup is later than those for other M-derived

clades restricted to Asia. Furthermore, the most ancestral M1 lineages are found in Northwest Africa and the Near East, but not in East Africa (Gonzalez et al., 2007). Therefore, the M1 haplogroup most likely originated from Asia, and expanded first to Northwestern Africa, then radiated to Eastern Africa and as far as the Iberian Peninsula. Remarkable parallelism can be observed between the M1 and U6 haplogroups in regard to the geographic distribution and time of origination. The analysis of a large number of complete mtDNA sequences of haplogroups M1 and U6 reveals that they both originated in Southwestern Asia and back-flowed together to Northwestern Africa through the Levant around 40,000 to 45,000 years ago, which temporally overlapped the settlement of Europe by modern humans (Olivieri et al., 2006). Interestingly, these early Upper Paleolithic lineages harboring haplogroups M1 and U6 traveled through the Mediterranean area instead of the original southern coastal route of the out-of-Africa expansion, possibly due to the improvement of climatic conditions that allowed for more permeable land.

Haplogroup H is a subclade of superhaplogroup R, under the HV branch, and is the most frequent haplogroup of North Africa as well as Western Eurasia, comprising almost half of the European mtDNA pool. Haplogroup H is present at 22% in the Near East, but only 9% in the Arabian Peninsula. Detailed molecular dissection of haplogroup H by sequencing of complete mtDNA has been carried out in a number of recent studies. This previously thought uniform haplogroup is refined to numerous monophyletic subclades, each bearing characteristic mutations and differentiated geographical distribution (Achilli et al., 2004; Loogvali et al., 2004; Pereira et al., 2005). The overall frequency of Haplogroup H in North Africa is highest in the western part, ranging between 24% to 37% of all mtDNA lineages in Moroccans, Algerians, and Tunisians, and drops slightly eastward to 14% to 21% in Egyptians and southward to 24% in Saharans (Rando et al., 1998; Krings et al., 1999;

Stevanovitch et al., 2004). Among subgroups of H, H1 has the highest frequency in North Africa at 42%, followed by H3 at 13%, both decreasing from west to east. This pattern of gradual change is similar to their frequency distribution observed in Europe, especially in the Iberian Peninsula. Most of other subgroups of H, H4, H5, H7, H8, and H11 have higher frequency in the eastern part of North Africa, attesting to possible gene flow from the Near East. The H1 and H3 subhaplogroups have a similar coalescent age of around 11,000 years in North Africa, indicating a late Paleolithic settlement (Ennaffa et al., 2009). The lack of exclusive haplotypic sharing between populations in North Africa and the Arabian Peninsula supports the hypothesis that the historical invasion and domination by Islamic Arabs has left strong influence in culture but only minor demic impact on North Africans (Bosch et al., 2000).

Binary and microsatellite genetic markers on the non-recombining region of Y chromosome provide another haploid haplotyping system to investigate human population evolution, although its molecular resolution is lower than that of mtDNA. A well-resolved phylogenetic tree of Y chromosome binary markers indicates that the top two primary splits branch out the Africa-specific haplogroups A and B. Both of them have various subclades with distinct geographic distribution, reflecting the complex population demographic history of Africa (Underhill et al., 2000; Hammer et al., 2001; YCC, 2002). The other part of the Y chromosome tree consists of three subgroups, C, DE, and F-M89, which coalesce at CR-M168 (Underhill et al., 2001). Haplogroup DE is present in both Africa and Asia, while haplogroup C is widely distributed in East Asia, Oceania, and North America, but not in Africa. Haplogroup F-M89 is also non-African and has a very prolific subgroup K with deep structure.

The most common Y chromosome haplogroup in North Africa is E3b2, at an overall frequency of 42%. It is a subclade under the DE branch, and present only at very low frequency in the immediate south of North Africa, the Near East, and

Southern Europe (Cruciani et al., 2004; Semino et al., 2004). The frequency of haplogroup E3b2 exhibits a decreasing pattern from west to east in North Africa, from 76% in the Saharawi of Morocco to only about 10% in Egypt (Bosch et al., 2001; Flores et al., 2001; Arredi et al., 2004). Haplogroup J* is the second most frequent in North Africa at an overall frequency of 20%, while being the most common among Palestinian Arabs and Bedouins. Weak negative selection has been proposed for E3b2 and J* haplogroups due to partial deletion of genes involved in the spermatogenesis, which should decrease their frequency (Repping et al., 2003), while no evidence for their positive selection has been reported so far. Therefore, the most likely explanation for the high frequency of E3b2 and J* haplogroups is neutral genetic drift. Given the possible weak negative selection on these haplogroups, the effective population size of male ancestors of North Africans is likely very small so that dramatic neutral drift could have happened. The time to most recent ancestor (TMRCA) for haplogroup E3b2 is estimated to be between 4,200 to 6,900 years ago, and between 6,800 to 7,900 years ago for haplogroup J*, modeled under various demographic parameters (Arredi et al., 2004).

Such young coalescence age of these two haplogroups further supports the hypothesis that neutral genetic drift is the predominant driving force shaping the genetic variation landscape of North Africans, at least on Y chromosome in males. The strong geographical structure of Y chromosomal variation and the parallel genetic affinity pattern in North Africa are consistent with the hypothesis that North Africans arise from a quick population expansion from the Middle East.

1.3 Population Genetics of Basques and Iberians

Basques are an ethic group living in the western Pyrenees in the Iberian Peninsula, traditionally known as the "Basque Country," across the border between northeastern Spain and southwestern France. Basques are generally considered to be a linguistically isolated population, as they speak the only non-Indo-European language, Euskera, in the Western Europe. Another unique feature of Basques is that they have the highest frequency of Rh negative blood type in the world around 33%, and almost no B or the related AB blood types. In addition, Basques exhibit particular distribution in the HLA system, immunoglobulin, and enzyme isoforms. It is also known that Basques have subtle but distinctive body characteristics, making them a distinct physical group. These peculiarities of Basques have been attributed to the long history of geographical isolation from the rest of Europe, and dramatic genetic drift due to small effective population size. Based on these particular biological features, it has long been thought that Basques are a genetically isolated population, and they might be the population most closely related to the Paleolithic ancestor of Europeans (Mourant, 1947; Aguirre et al., 1991; Bertranpetit and Cavalli-Sforza, 1991; Bauduer et al., 2005). In early population genetics studies, contemporary Basques often have been used as the proxy for the ancestral modern humans first settling in Europe, but this is no longer a useful paradigm.

Since the advent of molecular biology, DNA markers greatly facilitated population genetics research. Studies using various DNA genetic markers, including microsatellite DNA markers, mitochondrial DNA,Y chromosome markers, SNP markers in candidate genes as well as in the whole genome, provided further insight into Basque population genetics. Many genetic studies report evidence supporting the idea that modern Basques constitute a genetically distinctive population, while other

studies come to the opposite conclusion. Microsatellite DNA marker studies demonstrate that Basques can only be vaguely differentiated from neighboring populations in Europe and North Africa (Iriondo et al., 1997 and 2003; Belle et al., 2006; Zlojutro et al., 2006).

Analyses using haploid genetic markers reveal various levels of genetic differentiation of Basques in relation to surrounding populations. Mitochondrial DNA marker studies tend to demonstrate smaller degrees of genetic differentiation for Basques (Bertranpetit et al., 1995; Salas et al., 1998), while some mtDNA markers exhibit noticeable difference, such as a higher level of haplogroup H in Basques (Achilli et al., 2004). Y chromosome markers of Basques are found to be of conspicuously low genetic diversity, and share with the rest of European populations the most common haplogroup and the associated modal microsatellite haplotype (Seminoet al., 2000; Alonso et al., 2005; Adams et al., 2008). The Y chromosome genetic variation of Basques overall falls within the landscape of European genetic diversity, although a low degree of differentiation is proposed in some studies (Hurles et al., 1999; Rosser et al., 2000). Genetic variation studies using Alu insertions as well as SNP sets on selected autosomal segment are reported to fail to distinguish Basques from non-Basque Europeans (Comas et al., 2000; Garagnani et al., 2009).

With the advancement of DNA genotyping technology, especially the wide utilization of high-density DNA microarray, genome-wide SNP scans become feasible and affordable to apply in large-scale population studies that include large numbers of individuals. So far, the Human Genome Diversity Project (HGDP) is the genome-wide population genetics study that includes the largest number of populations worldwide, covering all major geographic regions, in which the French Basques are included (Li et al., 2008). In a report from the HGDP project, Basques turn out to be slightly distinguished from other European populations based on the genome-wide

SNP scan; however, it is unable to place Basques in a more refined geographical and genetic context among worldwide populations. Other high-density SNP genotyping studies have been carried out on Basques to compare them to other European populations (Rodriguez-Ezpeleta et al., 2010), as well as to other Iberian populations (Laayouni et al., 2010). These studies give rise to contradictory results in regard to the existence of genetic distinctiveness of the Basques. Compared to the genetic markers used in earlier studies, genome-wide SNP scans provide much more information about the evolutionary history of human populations. Therefore, further genome-wide genetic variation studies of Basques are necessary to resolve the controversy over their genetic characteristics.

CHAPTER 2

MATERIALS AND METHODS

2.1 Populations and Samples

The present study includes seven North African populations and four Spanish populations. Twenty individual samples were selected from each of these populations. Among North Africans, Egyptian samples were collected from Upper Egypt and different areas of the delta; Saharawi samples from the occidental Sahara; North Moroccan samples from Chefchaouen and Nador; South Moroccan samples from Erraschidia and Quarzazate; andrandom samples from the general populations of Algeria, Libya, and Tunisia. The Southern Spanish samples were collected in the Andalusia region; the Northwestern Spanish samples from the Galicia region; the Basque samples from the Basque Country; while the Canary Islands samples were collected from the Eastern Islands, the Western Island, and the general population. The vast majority of these individuals were males. Genomic DNA specimens were extracted from frozen peripheral blood using standard laboratory protocols and stored at –80°C.

2.2 Genotyping Methods

Affymetrix human SNP 6.0 genotyping arrays (Affymetrix, Santa Clara, CA) were utilized in the genotyping experiment. This microarray harbors 906,600 SNP markers covering 22 autosomes, both X and Y sex chromosomes, and mitochondrial

DNA. Genomic DNA samples were first examined by spectrophotometer for purity and by agarose gel electrophoresis for DNA integrity, to exclude degraded or impure samples. Five hundred nanograms of good-quality genomic DNA from each sample were used in the DNA labeling reaction, and hybridized to the SNP 6.0 array according to the standard protocol provided by the manufacturer. The arrays were scanned by an Affymetrix GeneChip Scanner 3000 7G, and array signal intensity CEL files were generated using Affymetrix GeneChip Operating System (GCOS) software. Quality control of the microarrays and the genotyping calls were generated using BirdSeed v2 algorithm in Affymetrix Genotyping Counsel 4.0 software.

2.3 Quality Assessment of Genotyping Microarray Data

A total of 205 SNP 6.0 arrays out of all 220 arrays passed the default ContrastQC threshold in the Genotyping Counsel 4.0 software, and the genotyping calls generated from them were exported as forward strand alleles. ContrastQC, the new QC metric based on each individual microarray, demonstrated good correlation with the actual genotyping call rate based on the whole set of microarrays (figure 1A). Microarrays with extremely low call rates tend to have exceptionally high overall heterozygocity rates (figure 1B), which is probably due to high level of genotyping error. In the 205 microarrays called, 4 microarrays had genotyping call rates less than 94% and high heterozygocity rates, so they were immediately excluded from further analysis. A small proportion of SNPs showed high genotyping missingness (figure 1C). A total of 817,325 SNPs out of all 906,600 SNPs passed the missingness cutoff of 0.05, in which 763,351 SNPs had minor allele frequency (MAF) of greater than 0.01 in the 201 arrays (figure 1D) and were used for further analysis.

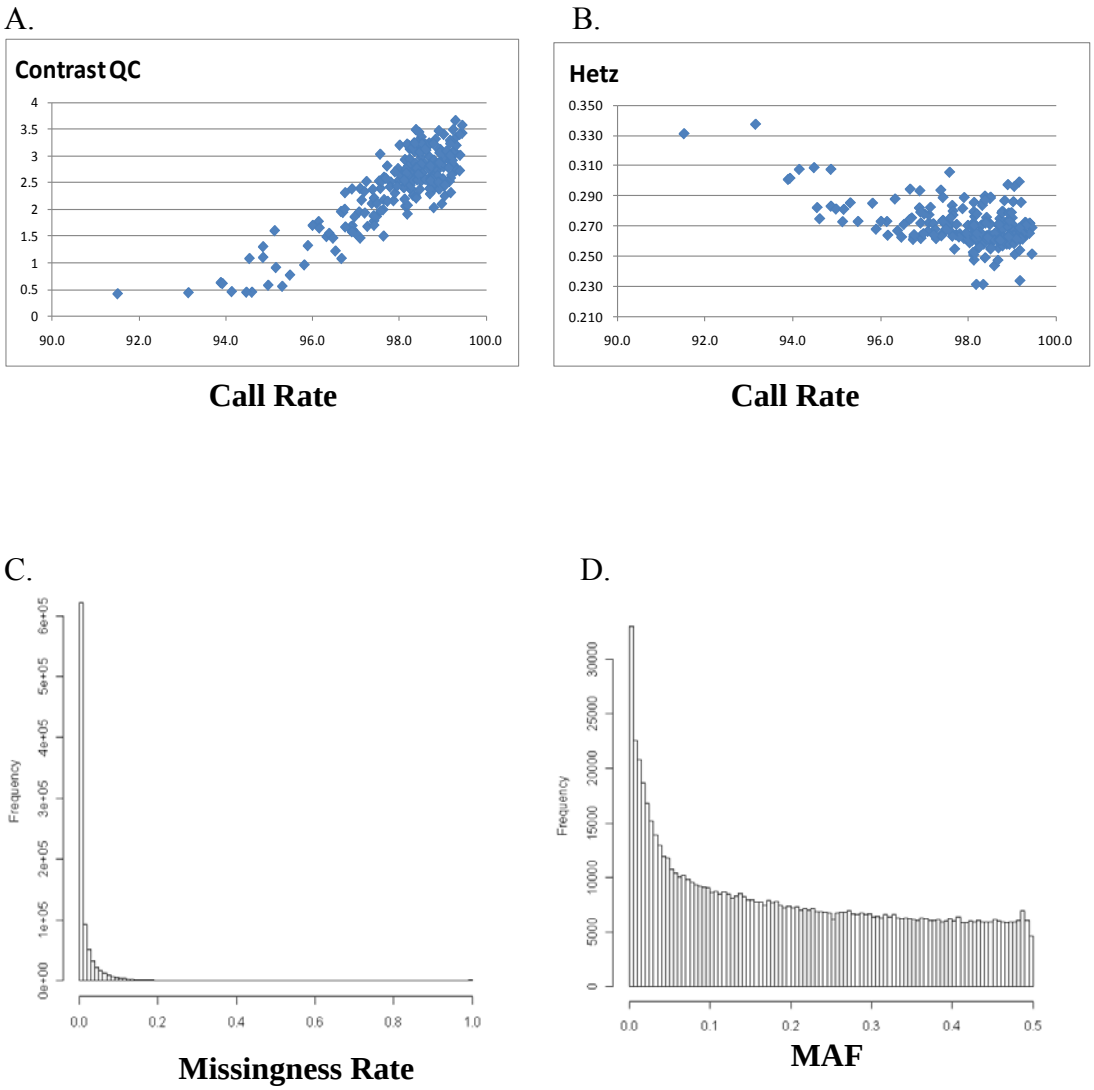


Figure 1. Quality control of genotyping data. A. Relationship between the microarray genotyping call rate and ContrastQC metric in all 205 arrays called. B. Relationship between the microarray genotyping call rate and the heterozygosity rate in all 205 arrays called. C. Distribution of SNP missingness rate in 201 arrays with call rate greater than 94%. D. Distribution of minor allele frequency (MAF) of 817,325 SNPs in 201 arrays with call rate greater than 94%.

SNP markers are expected to be in Hardey-Weinburg Equilibrium (HWE) in a population. Severe deviation from HWE is likely due to artifacts in generating SNP genotypes. The relationship among HWE p value, MAF, and observed heterozygosity rate of 817,325 SNPs were investigated (figure 2) using PLINK 1.07 software. SNPs very significantly deviated from HWE, such as with p values of less than 10^{-20} that demonstrated exceptionally high levels of MAF, mostly between 0.4 and 0.5 (figure 2A). Correspondingly, the heterozygosity rate of these SNPs were mostly between 0.8 and 1.0 (figure 2B). Such extremely high levels of the heterozygosity rate close to 1.0 is obviously unrealistic, and deviates significantly from the expected curve between MAF and the heterozygosity rate (figure 2C). These summary statistics are derived from the whole set of all 201 samples from 11 populations, hence the heterozygosity rate of these SNPs is also close to 1.0 in most of the individual populations. The most likely cause of such a high heterozygosity rate is a combination of poor performance of array probes detecting these SNPs and the bias of genotype calling algorithm to overcall hereozygotic genotypes on ambiguous SNPs, rather than due to biological or genetic reasons such as population stratification or isolation.

HWE p value is commonly used as a QC metric for filtering SNPs, but it is hard to set a threshold for the HWE p value in a practical and intuitive way, as it depends on the samples size and potential population stratification. By looking into the relationship between MAF and observed heterozygosity rate, it might be more useful to set a threshold directly on heterozygosity rate, instead of on the HWE p value, to filter out outlier SNPs of poor detection performance (figure 2B). Under expected HWE, the heterozygosity rate has a clearly defined relationship to MAF and a maximum of 0.5; any outlier heterozygosity rates much larger than 0.5 are likely artifacts. Meanwhile, the HWE p value should still be taken into consideration on a theoretical basis when establishing the cutoff on heterozygosity rate, as it exhibits

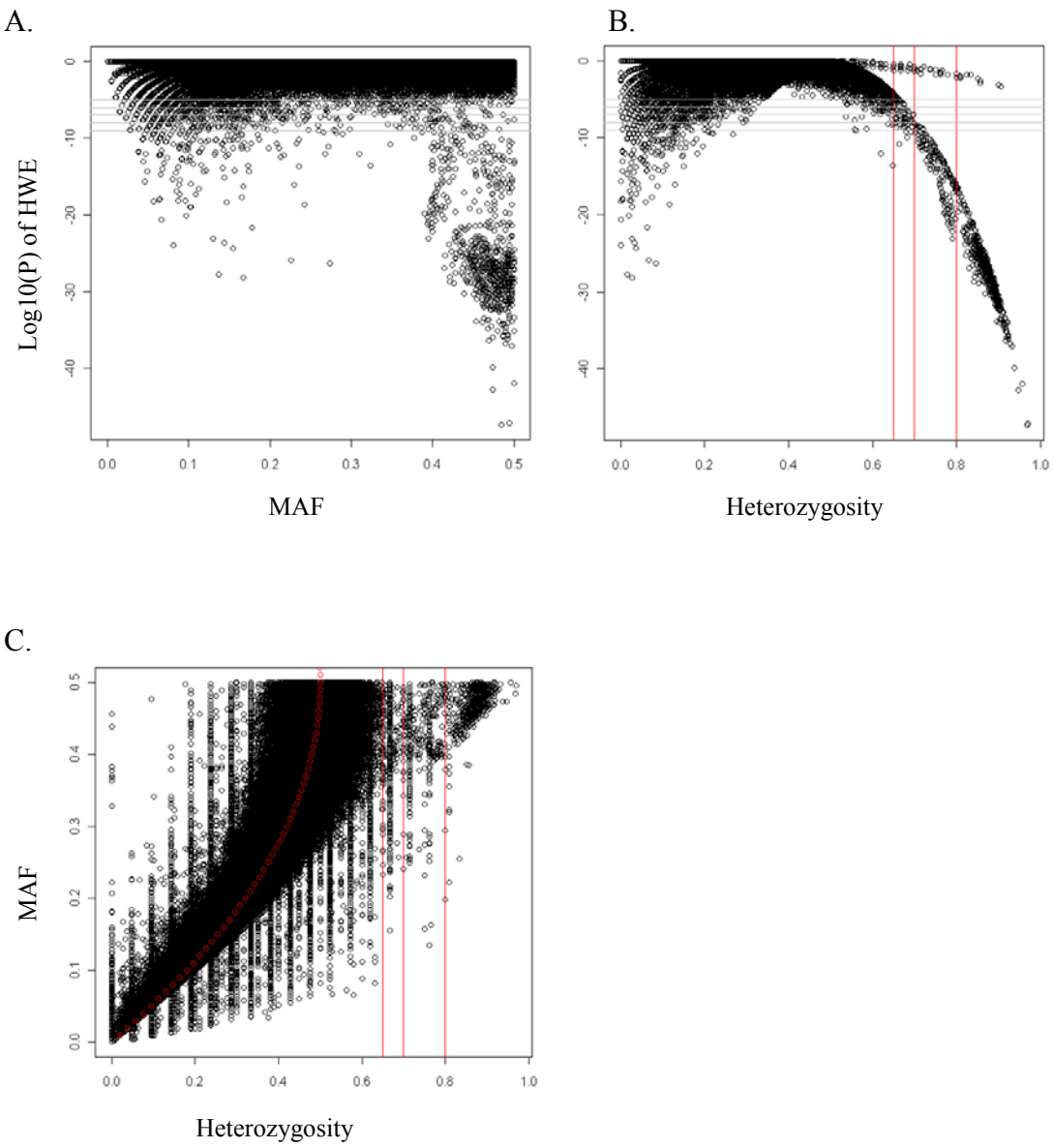


Figure 2. Pairwise relationship among SNP Hardey-Weinburg Equilibrium (HWE) p value, minor allele frequency (MAF), and heterozygosity rate. Each black circle represents an SNP, while the gray circles in (C) indicated the expected HWE.

a near linear relationship with high heterozygosity rate, such as greater than 0.65 in the present study (figure 2B).

Interestingly, a small set of SNPs turn out to have a high heterozygosity rate greater than 0.65 but an insignificant HWE p value of greater than 10^{-5} (at the upper-right-corner in figure 2B). Examination of their genome location reveals that these SNPs are all on the X chromosome. As most of the individuals studied are male, these heterozygous haploid genotypes of X chromosome SNPs in males are set to missing. The HWE p value is calculated only on the small number of nonmissing genotypes, and hence are insignificant. This actually illustrates one potential advantage of directly using the heterozygosity rate to filter SNPs, since these SNPs will be able to pass the HWE p value filtering but can be excluded due to high heterozygosity rate.

In the set of 817,325 SNPs examined in 201 samples, 1,459 SNPs with an observed heterozygosity rate of greater than 0.65 are excluded from further analysis. Only a maximum threshold is implemented on the heterozygosity rate in this data set, without a minimum threshold, to avoid overfiltering. The data set of the present study includes 11 populations of the small sample size of 20. The SNPs with an extremely low heterozygosity rate near 0 turn out to also have a very low MAF (figure 2C). These SNPs might be the most informative ones that are highly differentiated in one or a few of the populations and should not be excluded. Under other situations, a minimum threshold can be applied on the heterozygosity rate for appropriate SNP filtering. Actually, this reflects the flexibility of using the heterozygosity rate to filter SNPs; the cutoff can be set at the high end only or at both high and low ends. On the contrary, the HWE p value can have only one threshold, which might have different effects on SNPs of extremely high or low heterozygosity rates.

CHAPTER 3

RESULTS:

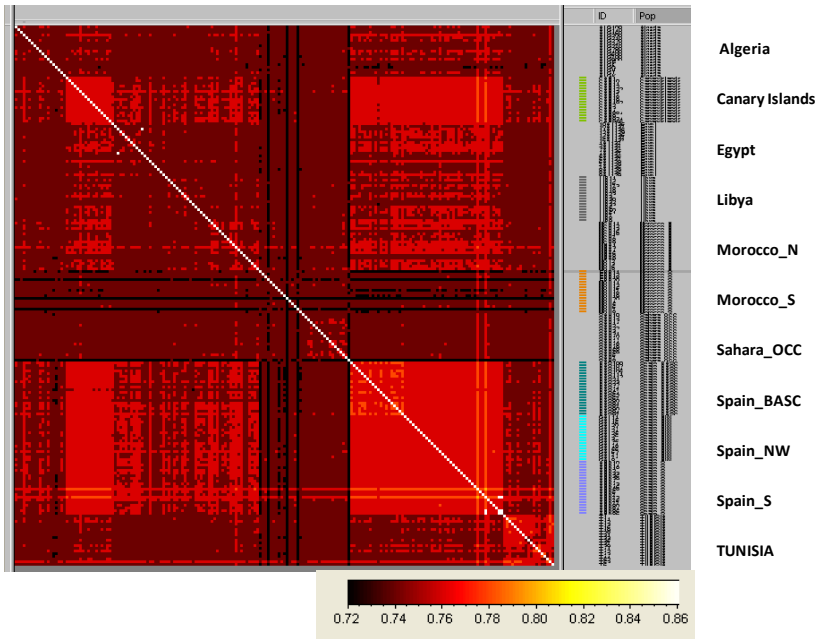
GENETIC VARIATIONS OF NORTH AFRICAN AND SPANISH POPULATIONS

3.1 Overall Identical-by-State Similarity Among 11 Populations

After filtering by SNP missingness, heterozygosity rate, and MAF, 762,587 SNPs were used to calculate the pairwise identical-by-state (IBS) matrix of 201 samples. Between each pair of individuals, there would be 0, 1, or 2 alleles shared at each SNP locus, and this could be averaged across all SNPs to obtain the mean proportion of IBS similarity between the individual pairs. The IBS matrix is a square matrix summarizing the proportion of IBS similarity of all possible pairs of individuals in the data set. The IBS matrix of 201 samples from seven North African populations and four Spanish populations are sorted by population and plotted with a color scale (figure 3A).

A few problematic samples were observed in the IBS matrix. One pair of Spanish samples were found out to be duplicates and showed a similarity of 0.997123, indicating high reproducibility in the genotyping experiment. Two Egyptian samples demonstrated exceptionally high IBS of 0.8625 to each other, suggesting that they might be close relatives. In these two pairs of related samples, only the ones with higher genotyping call rates were retained. In addition, two Spanish samples and one Tunisian sample showed high IBS and high heterozygosity rate compared to all other samples. They were all excluded due to potential DNA cross contamination.

A.



B.

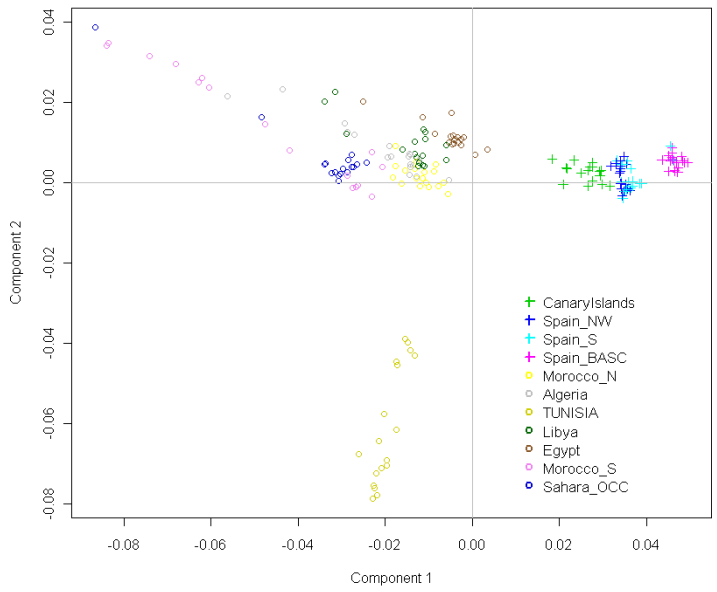


Figure 3. Pairwise IBS similarity matrix using the largest number of individuals and SNPs feasible: 201 individuals from 11 North African and Spanish populations with 763,351 SNPs with MAF > 0.01. A. IBS matrix were ordered by the population names. B. Multidimensional scaling (MDS) of IBS matrix. Top two components of MDS were plotted. probably due to potential DNA cross contamination; all three samples were, therefore, excluded.

Some interesting patterns of similarity among these 11 populations are revealed by the pairwise IBS matrix. The three Iberian and Canary Islands populations demonstrate higher IBS similarity within the group than they do to North African populations, while Basques have the highest within-population similarity. Surprisingly, most North African populations, including Algerians, Egyptians, Libyans, and North Moroccans, exhibit higher similarity to the Spanish populations than within the North African group. In North African populations, only the within-population similarity of Tunisians and South Moroccans is higher than their similarity to all other populations. However, Tunisians also show higher similarity to the Spanish populations than to other North Africans. Saharawi individuals show low similarity to all populations, including Saharawis themselves. Actually, they include most of the individuals with the lowest level of overall IBS similarity, which is visualized as black lines in figure 3A. These patterns suggest a complex and heterogeneous genetic architecture in North African populations.

An efficient way to summarize the variations among individuals in the pairwise IBS matrix is through multidimensional scaling (MDS). MDS is a statistical technique of dimensional reduction to partition the variance in the matrix into a series of orthogonal components ordered by the magnitude of variance explained. Applying MDS on the pairwise IBS matrix of 201 samples from 11 populations, the top two MDS components are plotted to explore the similarity and dissimilarity among individuals from different populations (figure 3B). MDS component 1, capturing the largest variance possible in one dimension, effectively separates Iberian populations from North African populations. In this dimension, the Southern Spanish and Northeastern Spanish populations overlap with each other, and are clearly separated from the Basque and Canary Islands populations. The Canary Islands population is closer to North Africans than the two Spanish populations are, while the Basque

population is more distant from North Africans than the two Spanish populations are. This is consistent with the demographic history of the Canary Islands. The Spanish colonization since the 15th century probably involved some admixture with indigenous populations, which might be related to North African populations. Basques are believed to be the population most closely related to the ancestors of Europeans, hence analysis including more European populations could shed additional light on the genetic characteristics of Basques. North African populations cannot be clearly separated from each other on MDS component 1, except that Egyptians form a tight cluster closest to the Spanish populations. South Moroccans encompass a wide range and include most of the individuals extending farthest away from the Spanish populations. MDS component 2 pulls out Tunisians far from the Spanish and other North African populations, with a few South Moroccan individuals being most distant from Tunisians.

MDS plot of IBS matrix reveals the heterogeneity within North African populations, especially in South Moroccans, Saharawi, and Tunisians. The sample sizes of the populations investigated are not large, ranging between 17 and 20; therefore, a few outlier samples in a population could substantially influence the characterization this population. Outlier samples might reflect the actual genetic heterogeneity in the population studied, it is also possible that they are caused by genotyping artifacts. Stringent filtering on the data quality of SNPs as well as of the samples could effectively reduce potential errors in subsequent analysis and inference of the data set. SNPs with missing genotype calls tend to be hard to detect accurately, and missingness in the genotype data set can be hard to handle or cause bias in computation. In the present study, 191 individuals have a genotyping missingness rate of less than 3% and are retained for further analysis. In these individuals, a total of

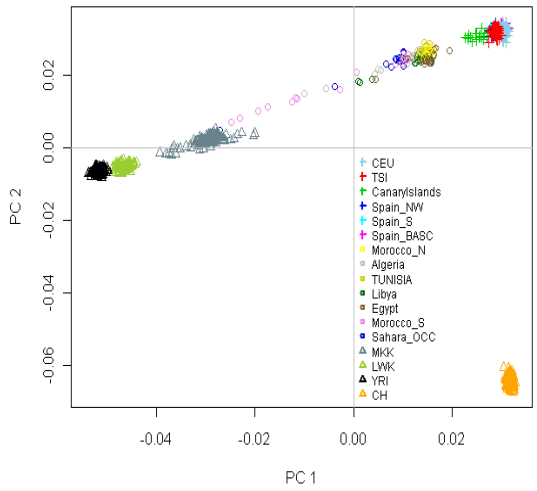
396,750 SNPs have complete genotype calls without any missing data and are used in subsequent analysis.

3.2 Principle Component Analysis Involving Other Representative Worldwide Populations

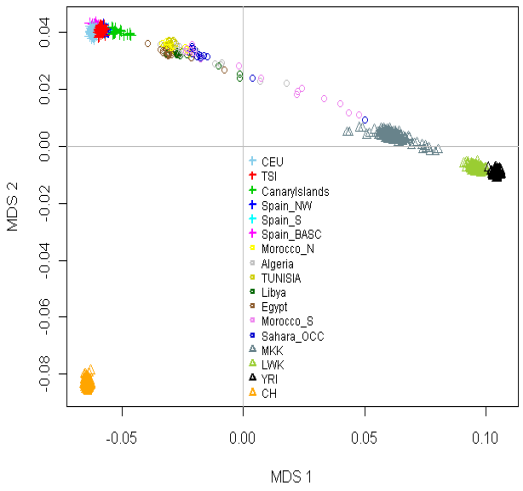
The 11 North African and Spanish populations in the present study wwere chosen to be located at the junction between the Eurasian and African continents. Therefore, it would be necessary to analyze these populations together with other representative worldwide populations to fully characterize the genetic features of these populations in the context of the whole world. There are two major public resources for genotyping data of worldwide populations, the Human Genome Diversity Project, which used only the Illumina BeadArray SNP genotyping platform (Cann et al., 2002; Rothenberg et al., 2002), and the International HapMap Project, which used both Illumina BeadArray and Affymetrix GeneChip platforms (International HapMap Consortium, 2007; 2003). The SNP content on Illumina BeadArray platforms and Affymetrix GeneChip platforms have a very small overlap of approximately 10%. As the present study used only the Affymetrix SNP 6.0 GeneChip, genotype data of populations from the latest HapMap phase 3 were selected to be combined with the present study, so that the maximum possible SNP content could be retained in the merged data set for further analysis.

Principal component analysis (PCA) on genotype data has been widely used to detect population structure on a genome-wode scale and to account for it in trait association studies (Price et al., 2007). The concept behind PCA is similar to MDS.

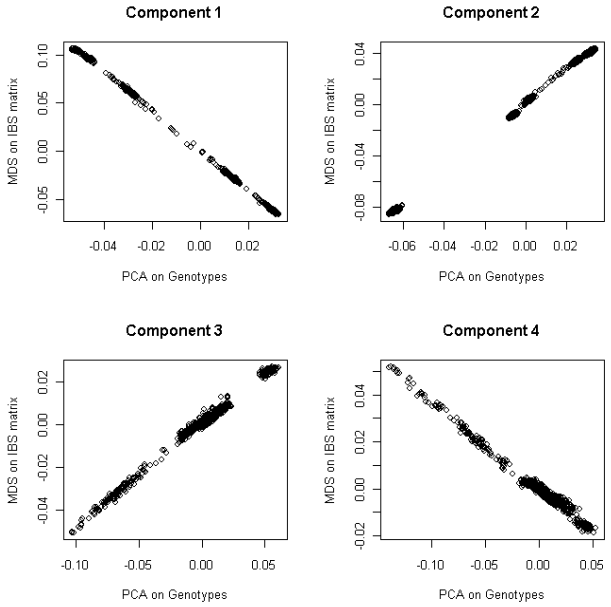
A.



B.



C.



D.

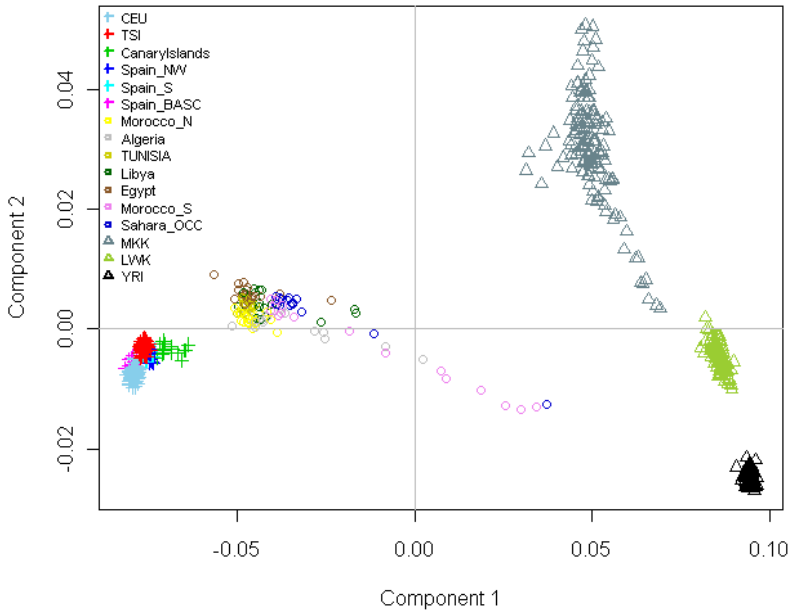
Population	Size	Group Size
CEU	106	258
TSI	84	
CanaryIslands	17	
Spain_NW	17	
Spain_S	17	
Spain_BASC	17	
Morocco_N	18	123
Algeria	18	
TUNISIA	18	
Libya	17	
Egypt	19	
Morocco_S	16	
Sahara_OCC	17	
MKK	138	333
LWK	85	
YRI	110	
CH	159	159

Figure 4. Population structure of 11 North African and Spanish populations merged with 7 HapMap3 populations (682 individuals) using 203,564 SNPs. A. Genotype Principal Component Analysis. B. MDS on IBS matrix. C. Relationship between PCA and MDS on the first 4 components. D. Size of each population included and 4 major groups of population.

The difference between them is that PCA is applied directly on the large genotype data matrix containing all the genetic markers status in all the individuals studied, while MDS is usually applied on the much smaller square matrix of pairwise IBS, with size equal to the number of individuals studied. In addition, the PCA requires genetic markers to be independent for matrix computation, therefore, SNPs in linkage disequilibrium (LD) need to be thinned before applying the PCA. LD in SNPs examined is not an issue for MDS on IBS square matrix, as all SNPs have already been summarized to the proportion of IBS sharing between each pair of individuals, although SNPs in high LD could increase the weight for these genomic regions and cause bias. It will be interesting to compare the result of the PCA on genotype matrix with MDS on the corresponding IBS matrix.

At a stringent LD r^2 threshold of 0.5, 220,346 SNPs out of the 396,750 SNPs without missing data in 191 North African and Spanish individuals were selected to merge with seven populations from HapMap3 data. These include three Sub-Saharan African populations, YRI, HWK, and MKK; two European populations, CEU, and TSI;two Chinese populations CHB and CHD, combined and named CH as they are indistinguishable from one another. A total of 203,564 SNPs out of the 220,346 SNPs were successfully merged between 191 samples of the present study and 682 samples of HapMap3 populations (figure 4D). Both genotype PCA (figure 4A) using SmartPCA software in the EIGNSTRAT package (Price et al., 2006) and MDS on the IBS matrix (figure 4B) using PLINK 1.07 software (Purcell et al., 2007) were applied on the same merged data set to explore the relationship of North African and Spanish populations among major worldwide populations. The two-dimensional reduction methods give very similar results in population structure: the corresponding components between the two methods turned out to be highly correlated to one another for the top four components examined, while the signs of components are

A.



B.

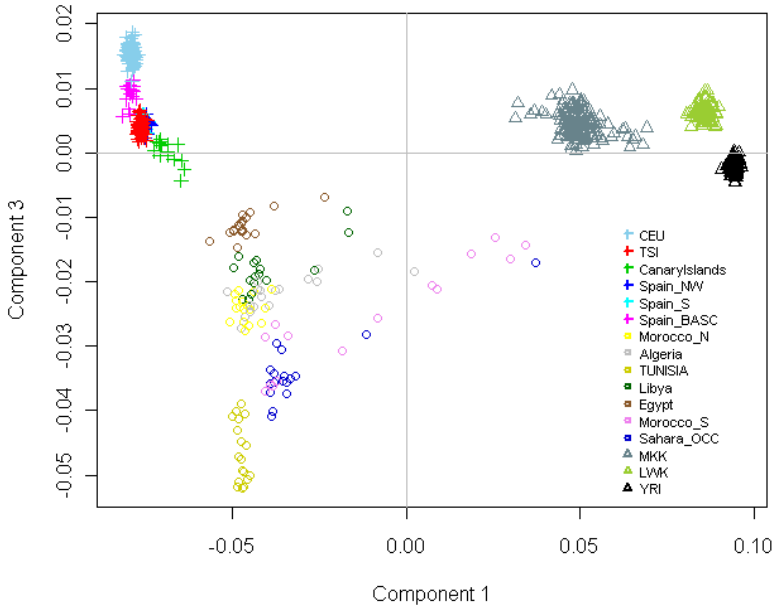


Figure 5. MDS plot on IBS matrix of 11 North African and Spanish populations merged with of 5 HapMap3 populations using 203,564 LD-pruned SNPs.

sometimes swapped (figure 4C). This indicates that the pairwise IBS matrix characterizing the inter-individual similarity proportion is able to retain almost all of the genotypic variance among individuals based on large numbers of SNPs. The computational load of MDS on IBS matrix is much lower than PCA on the large genotype data set, therefore MDS on IBS matrix provides a much more efficient computational method to explore the similarity patterns among individuals than genotype PCA.

Component 1 of both PCA and MDS separates Sub-Saharan African populations from European and Chinese populations, with YRI at the extreme. Among Sub-Saharans, the MKK population shows wide dispersion and is distant from YRI and LWK. Component 2 pulls the Chinese population away from all other populations, with Sub-Saharan African populations being closest to it, but still quite distant, and European populations at the opposite extreme. Looking at both components together, the four Spanish populations overlap with the European CEU and TSI populations, while the seven North African populations form a continuous line between Sub-Saharan Africans and Europeans. The Chinese population appears as a stand alone tight cluster far away from the continuous line formed by all other populations, hence, it is not informative in elucidating the genetic similarity of North African and Spanish populations and is excluded from further MDS analysis.

MDS is then applied on the IBS matrix of 714 individuals without Chinese, and the top three components are plotted (figure 5). Component 2 mostly differentiates MKK from the other two Sub-Saharan Africans YRI and LWK, exposing the large variation within MKK (figure 5A). Components 1 and 3 reveal the similarity pattern among populations at much better resolution (figure 5B). Fine structure among the European populations is observed. Three South European populations, including TSI, Northwest Spanish, and South Spanish, intermingle with

one another as a tight cluster. The CEU population of northwestern European origin extends away from the South European group, and it is the population most distant from Sub-Saharan and North African populations. Interestingly, Basques form a distinct cluster located right in the middle between the South European and the CEU populations. The Canary Islands population also extends away from the South European populations and toward North African populations. Tunisian is the North African population most distant from European and Sub-Saharan African populations on the axis of component 3, and has a wide range of variation within it. On this axis, the Tunisian population is followed by Saharawi and South Moroccan populations, while the Egyptian population is the closest to Europeans and Sub-Saharan Africans. MDS component 1 differentiates European populations from Sub-Saharan Africans, and North Africans turn out to be much more similar to Europeans than to Sub-Saharan Africans. However, a few North African individuals demonstrate a high level of similarity to Sub-Saharan Africans, especially the ones in Saharawi and South Moroccan populations that overlap with MKK individuals.

The pairwise IBS matrix of all 873 individuals, including HapMap3 Chinese, is averaged by population and visualized on a gray scale (figure 6). This is based on the 203,564 SNPs passing the stringent LD pruning r^2 threshold of 0.5, in order to prevent bias in overweighting genomic regions of extended LD. Basques stand out to be the population with highest similarity to the CEU. The similarity within North African populations appears to be lower than that within European populations, and mostly also lower than the similarity between the North African group and the European group. Saharawi and South Moroccan populations demonstrate low similarity to all European and North African populations, including within themselves. Individuals within the same population are expected to share more genetic similarity with one another than with other populations. Actually, this pattern is observed in all

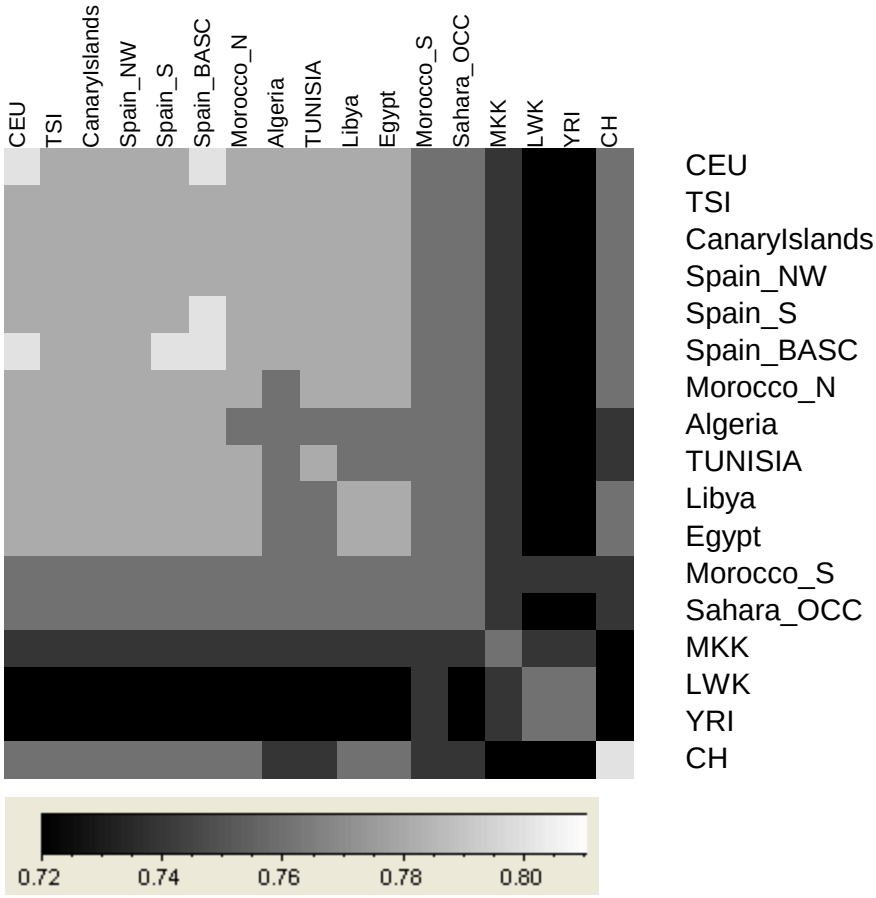


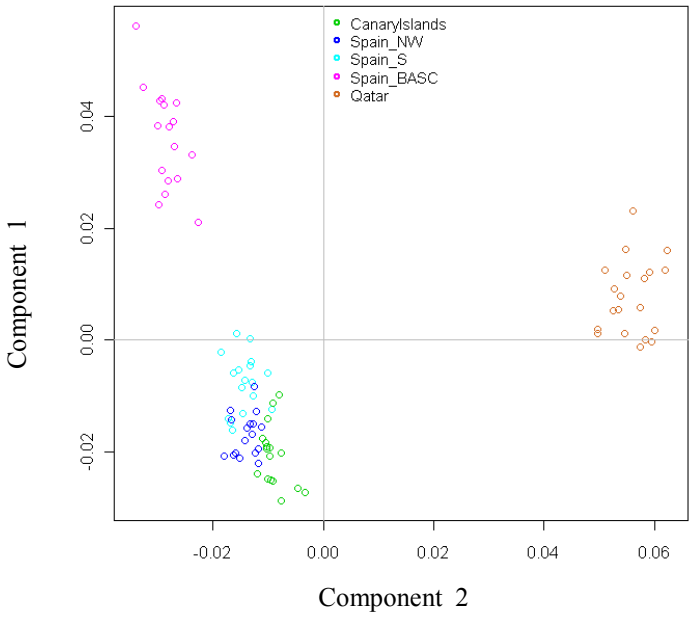
Figure 6. Population means of IBS matrix for each of the 11 North African and Spanish populations, as well as 6 HapMap3 populations, using 203,564 LD-pruned SNPs. The color scale is indicated at the bottom.

HapMap3 worldwide populations except the southern European TSI population. One possible cause for the absence of this pattern in the North African and Southern European populations might be the SNP ascertainment bias. The SNPs used in the large-scale genotyping assay are discovered mostly in the Western Europeans, as well as in Sub-Saharan African and East Asian individuals. Therefore, SNPs restricted to other populations that have not been thoroughly examined, such as in North African populations, would probably not be included in the SNP discoveries and subsequent genotype data set. Consequently, the corresponding genetic features might not be observed. More thorough, unbiased SNP discoveries in diverse populations will enable more powerful elucidation of distinct genetic characteristics in various populations around the world.

3.3 Genetic Distinctiveness of Basques Among Spanish Populations

Basques are generally considered a cultural isolate, speaking the only non-Indo-European language in the Western Europe. Many previous genetic studies reported that Basques constitute a genetically distinct population, using microsatellite DNA markers, genetic markers on Y chromosomes, mitochondrial DNA and candidate genes, and, lately, genome-wide SNP markers (Bertranpetit and Cavalli-Sforza 1991; Calafell and Bertranpetit 1994; Bertranpetit et al., 1995; Salas et al., 1998; Achilli et al., 2004; Zlojutro et al., 2006; Adams et al., 2008; Li et al., 2008; Garagnani et al., 2009). The present study uses high-density SNP microarrays to investigate genome-wide genetic variations of North African and Spanish populations, which include Basques and two surrounding populations at Northwest and South Spain. This provides a good opportunity to explore fine substructures of Spanish populations and possible genetic distinctiveness of Basques. Basques are considered

A.



B.

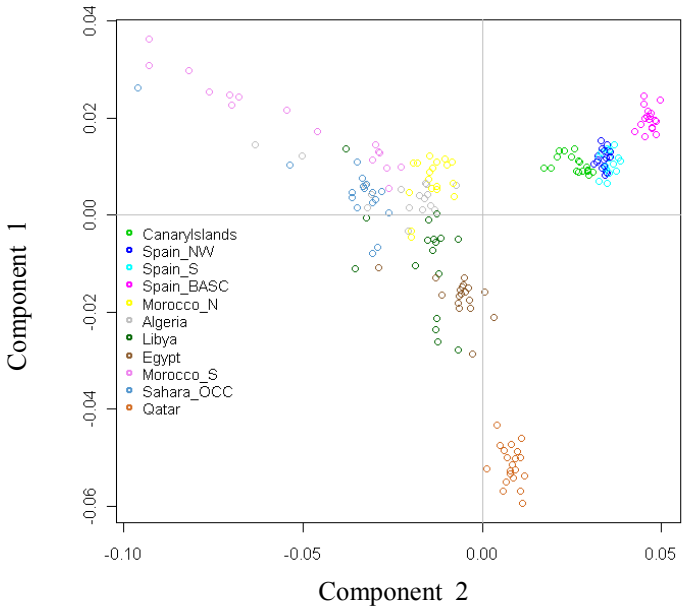


Figure 7. MDS plot of IBS matrix, including 20 individuals from the Qatari population, using 313,438 SNPs. A. MDS plot of Qatari merged with the four Spanish populations. B. MDS plot of Qatari merged with the 4 Spanish population and 7 North African populations.

the population most closely related to the ancestors of Europeans that probably have a Middle Eastern origin. Therefore, a set of 20 Qatari samples, genotyped on the Affymetrix Human SNP 500K platform, were obtained from collaborators (Clark et al., unpublished data) and merged with the data set of the present study.

The merged data set included 313,438 SNPs genotyped across 211 individuals from 12 populations. MDS plot was applied on the merged data set of Qatari and four Spanish populations, either with (figure 7B) or without (figure 7A) North African populations. Even when only Spanish and Qatar populations are analyzed together, the top two MDS components clearly separate Basques as a distinct cluster far away from the cluster formed by three other Spanish populations overlapping with each other, and away from the Qatari cluster (figure 7A). Component 1 indicates that Basques are more distant from Qatari than three other Spanish populations, while component 2 distinguishes Basques from the three other Spanish populations with Qatari in the middle between the two clusters. When analyzed together with North African populations, the fine substructure of Spanish populations is even more clearly revealed by the top two components of MDS on the IBS matrix (figure 7B). The first MDS component separates the Spanish populations from North Africans, with Qatari located between the two groups. The Northwest Spanish and South Spanish individuals are intermingled with one another, while Basques form a distinct cluster at one extreme end most distant from North Africans. Canary Islands individuals are also distinguishable from the Iberian populations on this axis, extending closer to North Africans and Qatari. The second MDS component is polarized by Qatari. Egyptians and Libyans are the North African populations closest to Qatari on both MDS axes, followed by Algerians and North Moroccans, then by Saharawi and South Moroccans. Interestingly, the observed pattern of relative genetic similarity among

North Africans and Qatari essentially corresponds to their geographical locations, similar to what has been reported about the fine genetic substructure of European populations (Novembre et al., 2008).

It is generally accepted that modern European populations originated from Southwestern Asia in Paleolithic times. With the introduction of the Middle Eastern Qatari population into the present analysis, the Basque and Canary Islands populations can be more effectively distinguished from surrounding populations of the Iberian Peninsula. The Qatari population also helps in the interpretation of genetic similarity pattern of North African populations by providing a means to polarize the change of genetic variations.

3.4 Runs of Extended Homozygosity Analysis

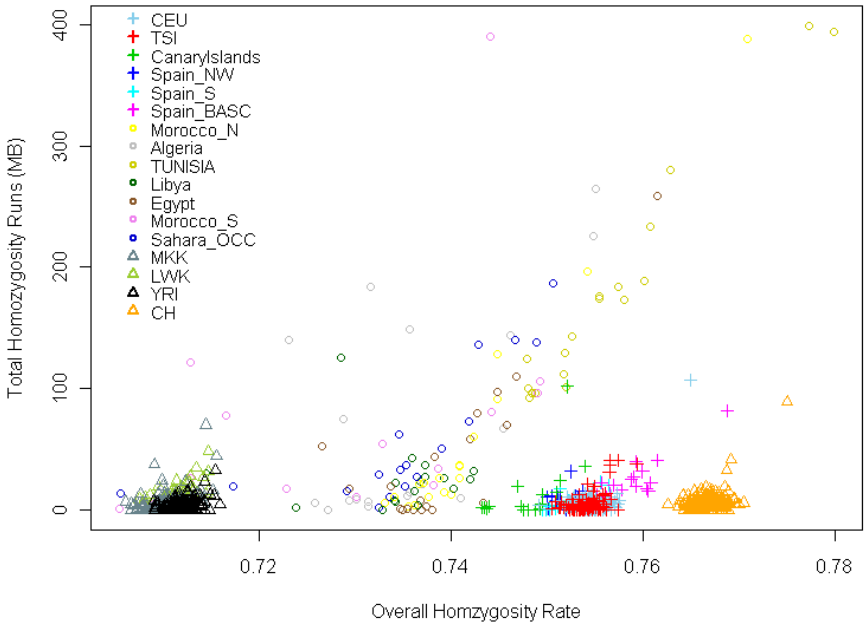
On a genome-wide level, extended tracts of homozygosity can be observed in an individual when both homologous chromosomal segments are inherited from a same recent common ancestor. Such runs of extended homozygosity (ROH) potentially can be the genomic indicator of autozygosity resulting from recent consanguinity (Li et al., 2006). In randomly mating populations, certain low levels of ROH can still be observed, which may be related to the demographics history of a population (Nalls et al., 2009; McQuillan et al., 2008). A smaller founder population or more severe growth bottleneck can lead to longer runs of homozygosity and larger numbers of runs. Therefore, runs of extended homozygosity are an important parameter to characterize a population.

Long runs of extended homozygosity were examined in the 873 individuals from seven North African populations, four Spanish populations, and six Hapmap3 populations, using PLINK 1.07 software (Purcell et al., 2007). ROH segments were

selected by the following criteria: in 5MB windows sliding across the whole genome, a given SNP will be called in ROH segments if more than 5% of all overlapping windows spanning the SNP are homozygous (allowing at most 5 missing genotypes and 1 heterozygous genotypes); the called ROH segments need to be at least 1MB long, containing at least 100 SNPs at a SNP density of no more than 50KB/SNP, with maximum gap between SNPs allowed at 1MB. The length of all ROH segments called in a given individual can be summed up to obtain the total length of ROH. Intuitively, the total length of homozygosity runs in an individual would be related to the overall SNP homozygosity rate, thus the relationship between the two was investigated (figure 8).

Looking at the population average (figure 8B), North African populations clearly demonstrate much larger total length of homozygosity runs than other populations, even though their overall SNP homozygosity rate is actually lower than European and Chinese populations. Tunisians have the highest level of total homozygosity runs, which is approximately three times as high as other North African populations and more than ten times asas high as European, Chinese, and Sub-Saharan populations. The overall SNP homozygosity rate of Tunisians is also the highest among North African populations, at a level comparable to most Europeans and only lower than Basque and Chinese. At the individual level (figure 8A), many North Africans exhibit exceptionally high levels of ROH, while some of the North African individuals have ROH level comparable to other populations. A linear trend between the total length of ROH and overall SNP homozygosity rate can be observed. Such high levels of autozygosity in North African populations is most likely due to the elevated level of recent consanguinity, which varies greatly among individuals and among populations.

A.



B.

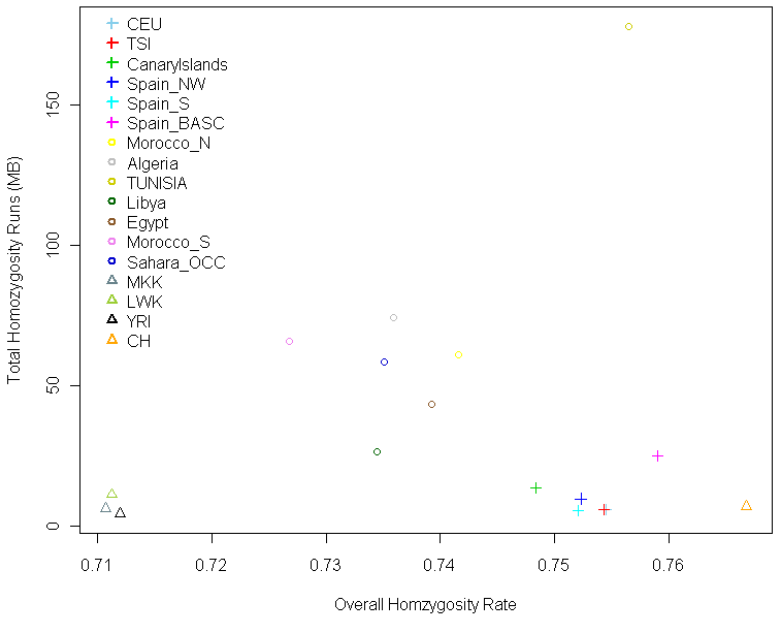


Figure 8. Relationship between the overall SNP homozygosity rate and total length of extended homozygosity runs. A. Plot of 873 individuals from North African, Spanish, and HapMap3 populations. B. Plot of population means.

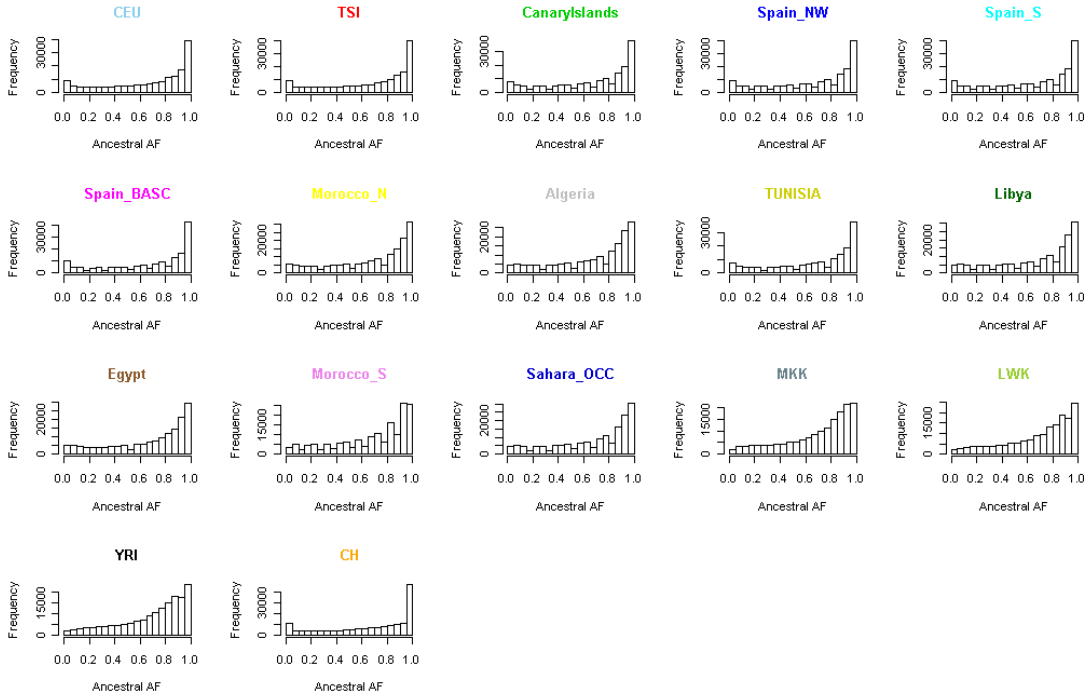
Among the four Spanish populations, the Northwest and South Spanish populations have comparable levels of autozygosity as the CEU and TSI, while Basques have levels approximately four times as high. The overall SNP homozygosity rate of Basques is also higher than any other European populations. This is consistent with the historical demographics of Basques. Basques are subjected to geographical isolation, which usually lead to elevated levels of inbreeding compared to randomly mating populations. The resulting autozygosity is manifested as runs of extended homozygosity in the genome. The Canary Islands also exhibit a slightly higher level of ROH than other Europeans except Basques, but its overall SNP homozygosity rate is the lowest among European populations. This can be explained by the migration history of the Canary Islands. The Spanish colonization of this area, beginning about 600 years ago, probably involved some degree of admixture with the indigenous populations; this gene exchange in turn lowered the overall homozygosity rate. In addition, the Canary Islands are also geographically isolated. All these factors contribute to the increased autozygosity in the Canary Islands.

These ROH results are consistent with known demographic history, culture, and geographical features of the populations studied. They indicate that runs of extended homozygosity analysis is an effective method to characterize a population with respect to inbreeding.

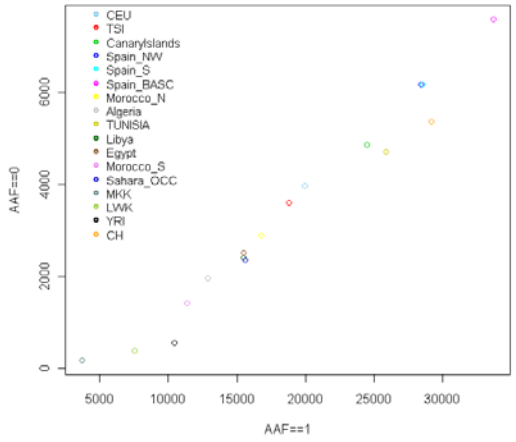
3.5 Ancestral Allele Frequency Characterization

Ancestral allele frequency (AAF) is an important parameter to characterize the evolutionary history of a population. The ancestral allele status information of SNPs assayed in the present study (Spencer et al., 2006) is obtained from NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/). A total of 167,888 SNPs out

A.



B.



C.

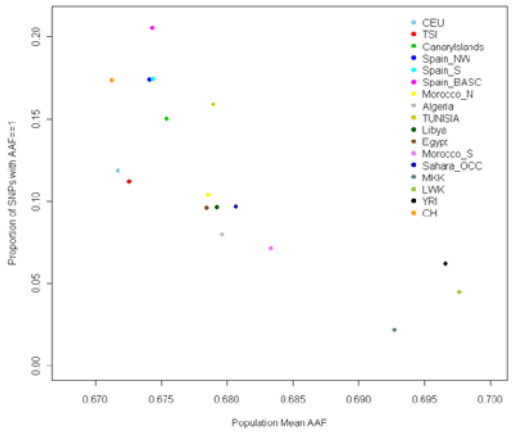


Figure 9. Ancestral allele frequency (AAF) characterization of 11 North African and Spanish populations as well as 6 HapMap3 populations. A. AAF distribution of each population. B. Relationship between the number of SNPs with fixed derived alleles and fixed ancestral alleles in each population. C. Relationship between the population mean AAF and the proportion of SNPs with fixed ancestral allele.

of the 203,564 LD-pruned SNPs have ancestral allele information available; their ancestral allele frequency distribution in each of the 17 populations is investigated (figure 9A). HapMap populations for Yoruban, European, and Chinese, demonstrate patterns similar to those previously reported (Li et al., 2008). Compared with other populations, Yorubans and other Sub-Saharan Africans have more SNPs at the upper end of ancestral allele frequency distribution and fewer SNPs at the lower end. Focusing on the intermediate range of AAF, excluding the two extreme ends, results in a steeper slope between the number of SNPs in each bin and the corresponding AAF. The Chinese appear to have a flatter slope than the CEU. The sample size of populations in the present study is much smaller than that of the HapMap project. This leads to fluctuated patterns in the ancestral allele frequency distribution, making it hard to obtain a reliable estimation of the slope between AAF and the number of SNPs in each bin.

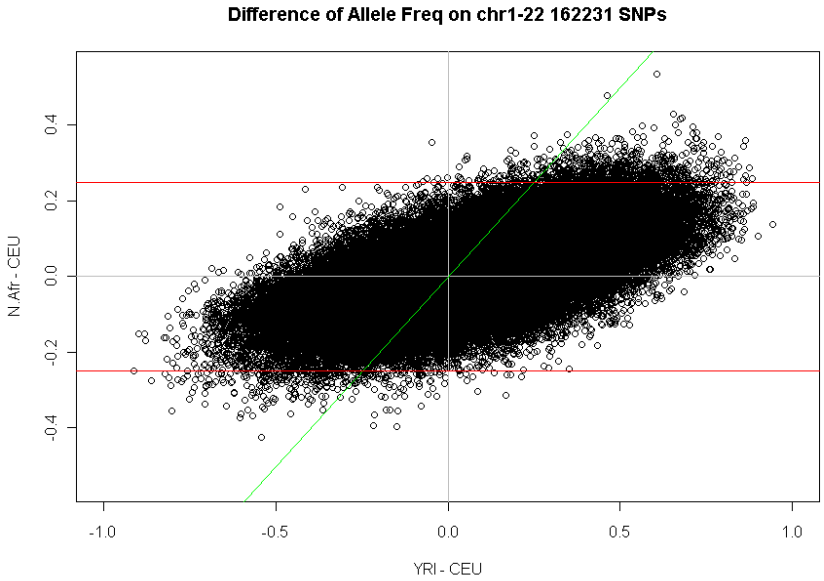
To circumvent this difficulty, SNP counts at the two extreme ends of AAF distribution, where AAF equals 0 or 1, is explored. A very strong positive linear correlation ($r^2 = 0.972$, slope = 0.257, p = 4.26e-13) is uncovered between the number of SNPs with fixed ancestral allele and those with fixed derived allele in each population (figure 9B). Surprisingly, it is the Basques who have the largest number of SNPs with fixed ancestral allele as well as with fixed derived allele among all populations, while the Tunisians are the highest among North African populations. Interestingly, a negative linear correlation (slope = –4.88, p = 4.12e-4, $r^2 = 0.61$) is observed between the population mean ancestral allele frequency and the number of SNPs with fixed ancestral allele in each population (figure 9C). Sub-Saharan Africans have the highest population mean AAF but smallest number of SNPs with fixed ancestral allele.

These observed patterns of ancestral allele frequency probably result from the combined effect of multiple demographic forces during the population evolution, and effective population size may play an important role in it. Populations having a smaller effective population size are more likely to undergo more dramatic genetic drift, to accumulate derived alleles more quickly to reach fixation, so they tend to have higher levels of fixed derived alleles and lower levels of genome-wide mean ancestral allele frequency. On the other hand, the newly derived alleles are also more likely to be lost as a result of the more pronounced genetic drift due to smaller effective population size, leaving the ancestral alleles to remain fixed in the population. Therefore, non-sub-Saharan African populations with a small effiective population size or severe bottleneck, such as North Africans and Europeans, have lower levels of overall ancestral allele frequency, but higher levels of both fixed ancestral alleles and fixed derived alleles.

3.6 Differential SNPs of North Africans

Ancestral allele frequency is an important parameter to characterize a population, hence, it is compared among populations in major geographical regions. The seven North African populations studied have a small sample size of 20 or less, and most of them are quite similar to one another, except the South Moroccan and Saharawi, who have several individuals admixed with sub-Saharan Africans. Hence, the ancestral allele frequency of the five North African populations, excluding South Moroccan and Saharawi, are averaged to obtain a more reliable estimate of AAF for this population group. Previous analysis results have shown that the genetic features of North African populations are much more similar to Europeans than to sub-Saharan Africans, therefore CEU is used as the baseline to characterize the AAF differential

A.



B.

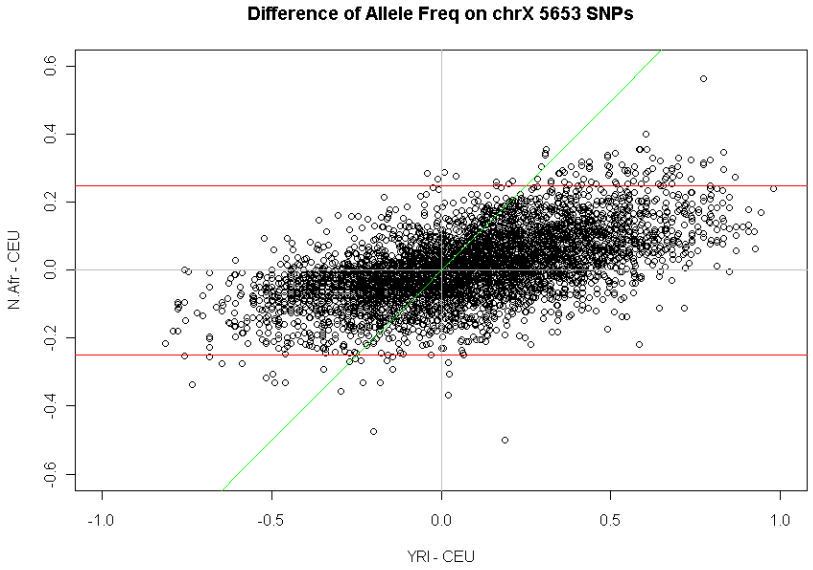


Figure 10. Ancestral allele frequency difference comparison of North Africans (not including Moroccans,_Southern, and Saharawis) versus CEU, and YRI versus CEU. A. Using 162,231 autosomal SNPs. B. Using 5,653 SNPs on X chromosome.

pattern of North Africans. The ancestral allele frequency difference of North Africans versus CEU is compared to the difference of YRI versus CEU on both autosomal SNPs (figure 10A) and SNPs on X chromosome (figure 10B). Positive correlation is observed in both SNP sets, the correlation coefficient is 0.64 for autosomal SNPs and 0.597 for SNPs on X chromosome. Noticeably, the regression slope was much less than 1. For SNPs with the most extreme difference in the North African versus CEU comparison, such as with absolute difference greater than 0.25, the direction of AAF change is almost all the same as YRI-CEU, but the magnitude of difference is much smaller than YRI-CEU. This differential pattern is consistent across the whole genome, indicating that North Africans have ancestral allele frequency mostly between sub-Saharan Africans and Europeans, with higher similarity to Europeans than to sub-Saharan Africans. SNPs having large AAF difference between North Africans and Europeans mostly exhibit even larger AAF difference between sub-Saharan Africans and Europeans.

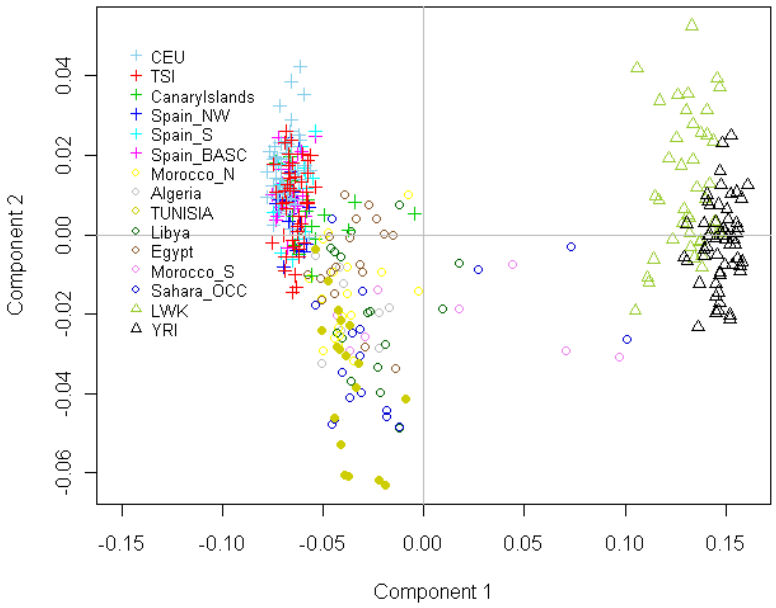
3.7 Analysis of X Chromosome in Males

The sample selection of the present study has a special feature that almost all individuals are male (figure 11C). The X chromosome in male is hemizygous with explicit genotype phase information readily available. This provides a good opportunity to explore the genetic variation of the X chromosome in the male populations in the present study. It is well known that population structure estimated from SNPs on X chromosomes tends to be more extreme than that from autosomes (Schaffner, 2004; Vicoso and Charlesworth, 2006). This can be attributed to a higher degree of population differentiation on the X chromosome loci compared to autosomal loci. Multiple factors may contribute to this. Both dominant and recessive mutations

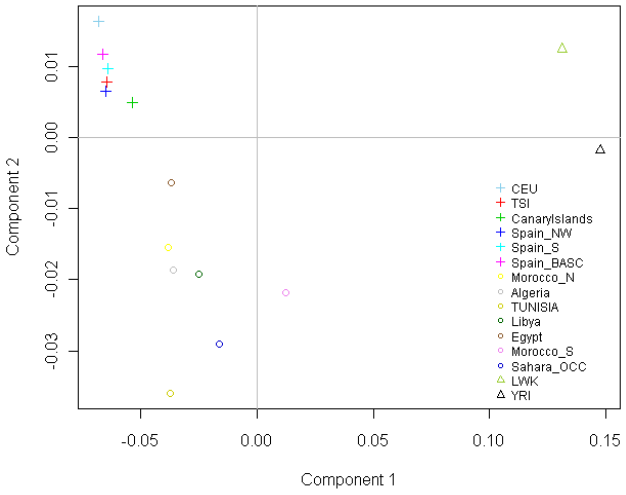
on the hemizygous X-linked loci are subjected to natural selection; the X chromosome has a smaller effective population size that leads to an elevated rate of genetic drift; there is also gender difference in reproduction, migration tendency, and generation time (Nielson et al., 2005; Hammer et al., 2008; Bustamente and Ramachandran, 2009; Keinan et al., 2009). Actually, it has been reported that the most differentiated SNPs among continental populations are significantly enriched on the X chromosome with a punctuated pattern (Charla et al., 2010).

A total of 5,653 SNPs out of the 203,564 LD-pruned SNPs are located on the X chromosome and have ancestral allele information available. Using these X-linked SNPs, MDS analysis was applied on the pairwise IBS matrix of 362 males from 11 North African and Spanish populations and four HapMap3 populations, CEU, TSI, YRI and LWK, to uncover population structure at better resolution (figure 11). The MKK and Chinese populations were not included, as they turned out to be uninformative in the previous analysis. The top two MDS components were plotted and revealed a pattern similar to what is observed from autosomal SNPs (figure 11). Component 1 separates Europeans from sub-Sahara Africans, while component 2 mostly separates North African populations from both Europeans and sub-Sahara Africans. However, one major difference between MDS patterns based on the two sets of SNPs is observed. Looking at the population average of MDS based on X chromosome, Tunisians are still clearly at one extreme end of component 2, but far less distinct from other North Africans than in the MDS based on autosomes (figure 5B). On the individual level, most of Tunisians intermingle with other North Africans in MDS on X chromosome, but they exhibit little overlap with other North Africans in MDS on autosomal SNPs.

A.



B.



C.

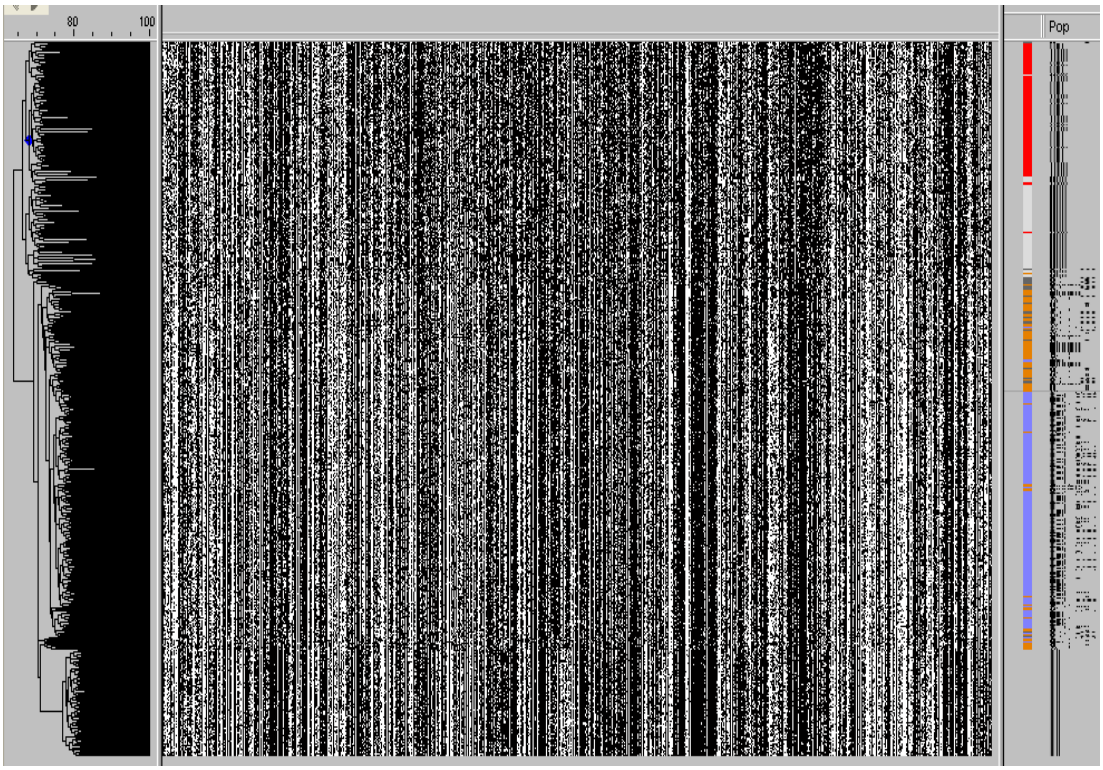
	Male	Female	Total
CEU	53	53	106
TSI	43	41	84
CanaryIslands	15	2	17
Spain_NW	17	0	17
Spain_S	17	0	17
Spain_BASC	17	0	17
Morocco_N	18	0	18
Algeria	8	10	18
TUNISIA	18	0	18
Libya	16	1	17
Egypt	19	0	19
Morocco_S	8	8	16
Sahara_OCC	17	0	17
MKK	67	71	138
LWK	42	43	85
YRI	54	56	110
CH	75	84	159

Figure 11. MDS plot on IBS matrix of 362 males from the 11 North African and Spanish populations, as well as 4 HapMap3 populations (CEU, TSI, YRI, LWK) using 5,653 SNPs on X chromosomes. A. MDS plot of individuals using top two components. Tunisians were represented as filled circles while others as are shown as open circles. B. Population means of top two MDS components. C. Population size by gender.

The pattern of population differentiation observed from MDS on the X chromosome in males appears to be opposite to the expected more extreme patterns. A high level of autozygosity is observed in North African populations, but not taken into consideration in the generally accepted hypothesis about the differentiation of the X chromosome. This may be the major cause of the differentiation pattern observed in the present study. Elevated autozygosity due to recent consanguinity could further increase the allele frequency difference between populations by reducing the effective population size of autosomal SNPs, but not the hemizygous SNPs of X chromosomes in males. Tunisians are the North African population most distant from Europeans and sub-Saharan Africans on the MDS plot, meanwhile, they also have the highest level of autozygosity. Therefore, the degree of genetic distinctiveness of Tunisians revealed by X chromosome SNPs in males is much lower than by autosomal SNPs. The autozygosity effect on SNP allele frequency difference between populations might make a substantial contribution to the excessively high level of population differentiation of Tunisians. In addition, the number of SNPs on the X chromosome is much smaller than that of autosomal SNPs. This may also lower the degree of population differentiation detectable using only X chromosomes.

Based on these X-linked 5,653 SNPs with explicitly phased haplotypes, the similarity pattern of these 362 males from 17 populations was explored using hierarchical cluster analysis (figure 12A). The SNP genotypes were represented by the number of ancestral allele, either 0 or 1, as columns in the dendrogram, ordered by their location on X chromosomes. Individuals were represented as rows in the dendrogram and clustered by Euclidean distance using complete linkage algorithm, with population groups indicated by colored sidebars. Individuals from different populations are mostly clustered by major geographical regions: European, sub-Saharan African, North African, and East Asian. However, fine population structure

A.



B.

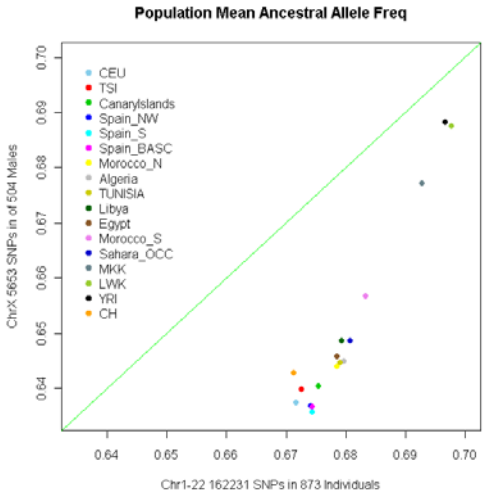


Figure 12. Analysis of 5,653 SNPs on the X chromosomes in 362 males from 17 populations. A. Cluster analysis. SNPs were plotted as columns ordered by genomic location, with ancestral allele in black and derived allele in white. Individuals were plotted as rows and clustered by Euclidean distance. Population groups were indicated by colored sidebars: North Africans by brown, except South Moroccans and Saharawis by dark gray, Europeans by blue, YRI and LWK by red, MKK by light gray, and CH by no color. B. Relationship between the population mean AAF of autosomes and X chromosomes.

within Europeans and North Africans cannot be clearly distinguished. Actually, European and North African populations appear to cluster next to each other and a small fraction of them are even intermingled with one another. South Moroccan and Saharawi are scattered among other North African populations, and some of them are clustered right next to sub-Saharan African populations. Chinese are distinct from other populations as a tight cluster with the highest within-population similarity, while sub-Saharan Africans exhibit the lowest levels of within-population similarity. Patterns of ancestral status in the hemizygous SNP haplotypes can be readily visualized in the dendrogram. Contiguous long tracks of same ancestral status, either ancestral or derived, are most obvious in Chinese populations, followed by Europeans and North Africans, while sub-Saharan Africans demonstrate mostly broken patterns of ancestral alleles intermingled with derived alleles.

Ancestral allele frequency (AAF) is an important parameter to characterize a population. Comparing the mean AAF of 5653 SNPs on X chromosomes in 504 males with that of 162,231 autosomal SNPs in all 873 individuals, a strong positive linear correlation ($r^2 = 0.95$) is uncovered between them (figure 12B). Sub-Saharan African populations have the highest mean AAF, while Europeans and Chinese are at the low extreme. North African populations lie between them, in much closer proximity to Europeans than to sub-Saharan Africans. Noticeably, the slope of the regression line is 2.03, with very significant p value of 3.88e-11. This indicates that the ancestral allele frequency difference between populations for SNPs on the X chromosome is on average approximately twice as high as that for autosomal SNPs. Therefore, X-linked SNPs do appear to have experienced more severe differentiation than autosomal SNPs during the population evolution, consistent with the commonly accepted hypothesis (Schaffner, 2004; Vicoso and Charlesworth, 2006).

In summary, the genome-wide SNP scan of North African and Spanish populations in the present study reveals that North Africans are genetically much more similar to the European and the Middle Eastern populations than to Sub-Saharan Africans. North Africans also demonstrate higher levels of homozygosity runs than other populations analyzed, probably due to consanguinity. Basques can be conspicuously distinguished from other Spanish populations and the Southern Europeans as having more genetic similarity to the Western Europeans.

CHAPTER 4

DISCUSSION

Previous population genetics studies of North Africans have been mainly based on small numbers of genetic markers of mitochondrial DNA, non-recombining region of Y chromosome, as well as microsatellite DNA. This study presents the first genome-wide scan of North African populations using high-density SNP markers. Uniparental haploid genetic markers have the obvious advantage of explicit haplotype phasing and sequential accumulation of mutations, making it straightforward to obtain a phylogenic hierarchy of the derived lineages (Underhill and Kivisild, 2007). However, these haploid genetic markers also have drawbacks and limitations, as they only constitute a very small portion of the whole human genome of each gender. Mitochondrial DNA is only 16,569 bp long, which can only harbor a limited number of mutations. This makes it hard to keep complete track of the highly stochastic process of population evolution. Y chromosomes harbor the longest non-recombining segment of DNA in the human genome, but the current molecular resolution, measured by SNP density, is much lower than in mitochondrial DNA. More SNP discovery effort on Y chromosome, such as deep re-sequencing, is needed to better reveal the genealogical architecture in males at higher resolution. Compared to the haploid genetic marker, the nuclear genome is subjected to recombination at various rates along the chromosomes and has low mutation rates, making it very complicated to build a genealogical tree from long sequences and hard to come up with a specific interpretation of the tree. However, the huge size of the nuclear genome allows it to retain a complete record of the human evolutionary history, including ancient genetic

drift events. High-density SNP scan of the whole nuclear genome can distinguish populations not only at the continental level (International HapMap Consortium, 2007), but also at the much finer regional level (Novembre et al., 2008; Auton et al., 2009; HUGO Pan-Asian SNP Consortium 2009). The phenotypic differentiation among human populations is genetically determined by the 3GB nuclear genome, and to a much smaller extent, by the tiny 16KB mitochondrial DNA. Therefore, analysis of genetic variations in the nuclear genome could potentially uncover the causal genetic components responsible for the phenotypic differentiation of populations.

The genome-wide SNP scan of North African populations reveals a clinal pattern of change in genetic similarity that corresponds to the geographical locations, as demonstrated by top components of MDS on pairwise IBS matrix. North Africans are genetically much more similar to Europeans than to sub-Saharan Africans, with the ratio of distance between the two populations at approximately 1:5 for most individuals (figure 5). However, a few outliers from different North African populations extend toward sub-Saharan Africans at varying degrees, forming a continuum between the majority of North Africans and the Maasai population in Kenya (MKK of HapMap3). This most likely results from recent admixture with sub-Saharans at different levels, as most of these outliers are South Moroccans and Saharawi at the Atlantic coastal end of the Sahara Desert who have more chances for contact with sub-Saharans. Such varying degrees of heterogeneity within each North African population, which has occurred naturally, will likely distort the population-level summary statistics estimation, such as SNP allele frequency, and lead to bias and complications in subsequent population genetics analysis. The best way to overcome this problem might be to increase the sample size so that outliers can be partitioned from the majority of the population studied and analyzed separately.

One of the top MDS axes is able to distinguish North African populations from both sub-Saharan Africans and Europeans. Tunisians are located at one extreme end of this axis opposite to Europeans and sub-Saharans; all other North African populations are scattered between, them roughly corresponding to their geographical locations. The South Moroccans and Saharawi exhibit the highest level of genetic similarity to Tunisians, and Egyptians the lowest genetic similarity to Tunisians. Therefore, this axis of variation may capture the autochthonous genetic component of North Africans. It is likely the Tunisian population contains the most indigenous Berber ancestry, while historically Egypt has been influenced more by the Southwestern Asia. When replacing Tunisians with the Middle Eastern Qatari in the MDS analysis, the relationship between genetic similarity and geographical location is more clearly observed in North Africans. Egyptians appear to be the most similar to Qatari, while the South Moroccans and Saharawi seem the least similar to Qatari (figure 7). These results are mostly consistent with what is uncovered from mitochondrial DNA and Y chromosome, and the hypothesis that North African populations originated from Southwestern Asia, probably through Levant during the same demic expansion that led to the settlement of modern humans in Europe (Olivieri et al., 2006). The genome-wide SNP scan in the present study clearly demonstrates that sub-Saharan Africans have only a small contribution to the genomic content of North African populations. In spite of high levels of genetic similarity, European populations, including Iberians, can be conspicuously distinguished from North Africans on the whole genome level. As to the geographical barriers of North Africa, the Strait of Gibraltar turns out to be more effective in preventing gene exchange, as revealed by the genome-wide similarity pattern, than the Sahara Desert, although cultural isolation may also contribute to this.

High-density SNP scan has also uncovered an interesting pattern of runs of extended homozygosity in North Africans. Overall, North Africans exhibit much larger total length of homozygosity runs than Europeans, sub-Saharans, and East Asians, probably due to elevated levels of consanguinity (figure 8). Noticeably, Tunisians have the highest level of total homozygosity runs, about three times as high as other North Africans and more than ten times higher than non–North African populations. The other two populations with the second highest level of indigenous Berber componet, the South Moroccans and Saharawi, also exhibit high levels of homozygosity runs among North African populations. Therefore, higher levels of autozygosity seem to correspond to higher levels of indigenous Berber components in North Africans. Consanguinity and endogamy are common in Berber as well as in the Islamic culture, especially in isolated rural areas (Arab et al., 2004). It is possible that both the cultural and geographical isolation, which may be associated with increased level of autozygosity, helped to preserve the ancient indigenous Berber ancestry through the present time.

It is generally accepted that SNPs on the X chromosome tend to exhibit more extreme population structure than autosomes due to more dramatic population differentiation (Schaffner, 2004; Vicoso and Charlesworth, 2006). It is observed in the present study that the ancestral allele frequency difference between populations for SNPs on X chromosomes is roughly twice as large as that for autosomal SNPs, confirming that X-linked SNPs do have more dramatic differentiation than autosomal SNPs. However, the extent of population differentiation detected by MDS on the X chromosome SNPs is much smaller than that on autosomal SNPs (figure 11). This appears to be contrary to the expected more extreme pattern on X-chromosome SNPs. The loss of resolution to distinguish populations is most likely due to the much smaller number of SNPs on X chromosomes than on autosomes.

Ancestral allele frequency is characterized for each North African population, and also as a whole group after excluding the South Moroccans and Saharawis, which have too many individuals admixed with sub-Saharan Africans. Compared to the Western European CEU population, the differential SNPs of the North African population group shows mostly the same direction of change in ancestral allele frequency as the sub-Saharan YRI population (figure 10). In fact, SNPs having large AAF difference between North Africans and Europeans mainly exhibit even larger AAF difference between sub-Saharan Africans and Europeans in the same direction. This differential pattern is consistent across the whole genome, similar for both autosomal SNPs and X-chromosome SNPs. This indicates that North Africans retain ancestral allele frequency at levels between sub-Saharan Africans and Europeans, and much closer to Europeans than to sub-Saharans. Interestingly, this pattern of genetic similarity corresponds well with the relative geographic location of these three large subcontinental regions from north to south. This may result from prehistorical human evolution or historical migration, or a combination of both.

The fine substructure of Spanish populations is also explored in the present study. The Canary Islands individuals are conspicuously separable from other Spanish populations in the genome-wide SNP scan. Compared to other Spanish populations, they exhibit more genetic similarity to the North Africans and sub-Saharans, as well as slightly lower levels of homozygosity rate but slightly more elevated levels of autozygosity. These genetic characters are in agreement with the known demographic history of the Canary Islands, such as recent Spanish colonization and associated admixture with the indigenous populations in this geographically isolated area. Admixture leads to elevated levels of genetic heterozygosity, while isolation tends to increase the level of inbreeding.

Basques are clearly distinguishable from other Iberian populations on the whole genome level, as being more similar to the Western Europeans represented by CEU (figure 5). This is consistent with previous reports on the existence of the genetic distinctiveness of Basques on the genome-wide level (Li et al., 2008; Rodriguez-Ezpeleta et al., 2010). One puzzling observation in genetic variation pattern is that Basques appear to be less similar to the Middle Eastern Qatari populations than other Spanish populations. Modern Europeans are believed to be originated from the Southwestern Asia, and Basques are widely thought to be the European population most closely related to their Middle Eastern ancestor. However, the contemporary Qatari population may not be a good proxy for the prehistoric ancestor of Europeans from the Southwestern Asia. This might be one major reason for the lack of greater genetic similarity between the Basque and Qatari populations. The other possible cause for this genetic similarity pattern may be the SNP ascertainment bias in the microarray platform used in the present study.

Additional population genetics features of Basques are uncovered in the present study. Basques have the highest level of autozygosity in European populations studied, a few times higher than the Southern and Western Europeans. On the other hand, Basques also have the largest number of SNPs with fixed ancestral allele as well as derived allele among all worldwide populations in the study, while their population mean ancestral allele frequency is the same as the other two Iberian populations (figure 9). The ancestral allele frequency distribution of the Basques is most similar to that of the East Asians represented by the Chinese, with a considerably large number of SNPs at both extreme ends, AFF = 1 and AAF = 1, yet a low and flat profile between them. According to the well-accepted out-of-Africa serial founder model of human expansion, this kind of ancestral allele frequency spectrum is indicative of high levels of genetic drift in the more peripheral populations with smaller effective

population size during the sequential chain of expansion (Li et al., 2008). Therefore, these ancestral allele frequency characteristics of Basques support the hypothesis that Basques have a small effective population size. However, the elevated level of autozygosity due to consanguinity caused by geographical and cultural isolation may also contribute to the formation of these genetic characteristics.

One almost inevitable difficulty in human population genetics studies is that human samples are available almost solely from the contemporary populations, although the latest advancements in technology make it possible to obtain genetic data from a few sources of ancient human DNA, such as an extinct Paleo-Eskimo individual (Gilbert et al., 2008; Rasmussen et al., 2010). The present-day residents in certain areas, including the indigenous inhabitants, may not directly represent the population living in the same area in ancient times. With the increasingly broader range and faster pace of globalization, even individuals from the indigenous populations will become harder and harder to have access to. As to Basques, too many demographic and evolutionary events could have happened during the long past, which could lead to various population genetics differences between Basques in the ancient time and ones in the present time. Therefore, the contemporary Basques probably cannot be simply considered as a "living fossil" to represent the ancestors of Europeans.

REFERENCES

Achilli, A., Rengo, C., Battaglia, V., Pala, M., Olivieri, A., Fornarino, S., et al. (2005). Saami and berbers--an unexpected mitochondrial DNA link. *American Journal of Human Genetics, 76*(5), 883-886.

Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R., 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. *American Journal of Human Genetics, 75*(5), 910-918.

Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R., 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. *American Journal of Human Genetics, 75*(5), 910-918.

Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R., 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. *American Journal of Human Genetics, 75*(5), 910-918.

Adams, S. M., Bosch, E., Balaresque, P. L., Ballereau, S. J., Lee, A. C., Arroyo, E., et al. (2008). The genetic legacy of religious diversity and intolerance: Paternal lineages of christians, jews, and muslims in the iberian peninsula. *American Journal of Human Genetics, 83*(6), 725-736.

Alonso, S., Flores, C., Cabrera, V., Alonso, A., Martin, P., Albarran, C., et al. (2005). The place of the basques in the european Y-chromosome diversity landscape. *European Journal of Human Genetics : EJHG, 13*(12), 1293-1302.

Arredi, B., Poloni, E. S., Paracchini, S., Zerjal, T., Fathallah, D. M., Makrelouf, M., et al. (2004). A predominantly neolithic origin for Y-chromosomal DNA variation in north africa. *American Journal of Human Genetics, 75*(2), 338-345.

Auton, A., Bryc, K., Boyko, A. R., Lohmueller, K. E., Novembre, J., Reynolds, A., et al. (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Research, 19*(5), 795-803.

Barbujani, G., Pilastro, A., De Domenico, S., & Renfrew, C. (1994). Genetic variation in north africa and eurasia: Neolithic demic diffusion vs. paleolithic colonisation. *American Journal of Physical Anthropology, 95*(2), 137-154.

Bauduer, F., Feingold, J., & Lacombe, D. (2005). The basques: Review of population genetics and mendelian disorders. *Human Biology; an International Record of Research, 77*(5), 619-637.

Belle, E. M., Landry, P. A., & Barbujani, G. (2006). Origins and evolution of the europeans' genome: Evidence from multiple microsatellite loci. *Proceedings.Biological Sciences / the Royal Society, 273*(1594), 1595-1602.

Ben Arab, S., Masmoudi, S., Beltaief, N., Hachicha, S., & Ayadi, H. (2004). Consanguinity and endogamy in northern tunisia and its impact on non-syndromic deafness. *Genetic Epidemiology, 27*(1), 74-79.

Bertranpetit, J., & Cavalli-Sforza, L. L. (1991). A genetic reconstruction of the history of the population of the iberian peninsula. *Annals of Human Genetics, 55*(Pt 1), 51-67.

Bertranpetit, J., & Cavalli-Sforza, L. L. (1991). A genetic reconstruction of the history of the population of the iberian peninsula. *Annals of Human Genetics, 55*(Pt 1), 51-67.

Bertranpetit, J., Sala, J., Calafell, F., Underhill, P. A., Moral, P., & Comas, D. (1995). Human mitochondrial DNA variation and the origin of basques. *Annals of Human Genetics, 59*(Pt 1), 63-81.

Bosch, E., Calafell, F., Comas, D., Oefner, P. J., Underhill, P. A., & Bertranpetit, J. (2001). High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern africa and the iberian peninsula. *American Journal of Human Genetics, 68*(4), 1019-1029.

Bosch, E., Calafell, F., Perez-Lezaun, A., Clarimon, J., Comas, D., Mateu, E., et al. (2000). Genetic structure of north-west africa revealed by STR analysis. *European Journal of Human Genetics : EJHG, 8*(5), 360-366.

Bosch, E., Calafell, F., Perez-Lezaun, A., Comas, D., Mateu, E., & Bertranpetit, J. (1997). Population history of north africa: Evidence from classical genetic markers. *Human Biology; an International Record of Research, 69*(3), 295-311.

Calafell, F., & Bertranpetit, J. (1994). Principal component analysis of gene frequencies and the origin of basques. *American Journal of Physical Anthropology, 93*(2), 201-215.

Cann, H. M., de Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., et al. (2002). A human genome diversity cell line panel. *Science (New York, N.Y.), 296*(5566), 261-262.

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature, 325*(6099), 31-36.

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research, 15*(11), 1496-1502.

Clark, J. D., Beyene, Y., WoldeGabriel, G., Hart, W. K., Renne, P. R., Gilbert, H., et al. (2003). Stratigraphic, chronological and behavioural contexts of pleistocene homo sapiens from middle awash, ethiopia. *Nature, 423*(6941), 747-752.

Comas, D., Calafell, F., Benchemsi, N., Helal, A., Lefranc, G., Stoneking, M., et al. (2000). Alu insertion polymorphisms in NW africa and the iberian peninsula: Evidence for a strong genetic boundary through the gibraltar straits. *Human Genetics, 107*(4), 312-319.

Corte-Real, H. B., Macaulay, V. A., Richards, M. B., Hariti, G., Issad, M. S., Cambon-Thomsen, A., et al. (1996). Genetic diversity in the iberian peninsula determined from mitochondrial sequence analysis. *Annals of Human Genetics, 60*(Pt 4), 331-350.

Cruciani, F., La Fratta, R., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., et al. (2004). Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of africa. *American Journal of Human Genetics, 74*(5), 1014-1022.

Day, M. H. (1969). Omo human skeletal remains. *Nature, 222*(5199), 1135-1138.

Ennafaa, H., Cabrera, V. M., Abu-Amero, K. K., Gonzalez, A. M., Amor, M. B., Bouhaha, R., et al. (2009). Mitochondrial DNA haplogroup H structure in north africa. *BMC Genetics, 10*, 8.

Excoffier, L. (2002). Human demographic history: Refining the recent african origin model. *Current Opinion in Genetics & Development, 12*(6), 675-682.

Flores, C., Maca-Meyer, N., Perez, J. A., Hernandez, M., & Cabrera, V. M. (2001). Y-chromosome differentiation in northwest africa. *Human Biology; an International Record of Research, 73*(4), 513-524.

Forster, P., & Matsumura, S. (2005). Evolution. did early humans go north or south? *Science (New York, N.Y.), 308*(5724), 965-966.

Garagnani, P., Laayouni, H., Gonzalez-Neira, A., Sikora, M., Luiselli, D., Bertranpetit, J., et al. (2009). Isolated populations as treasure troves in genetic

epidemiology: The case of the basques. *European Journal of Human Genetics : EJHG, 17*(11), 1490-1494.

Garcea, E. A., & Giraudi, C. (2006). Late quaternary human settlement patterning in the jebel gharbi. *Journal of Human Evolution, 51*(4), 411-421.

Garrigan, D., & Hammer, M. F. (2006). Reconstructing human origins in the genomic era. *Nature Reviews.Genetics, 7*(9), 669-680.

Gilbert, M. T., Kivisild, T., Gronnow, B., Andersen, P. K., Metspalu, E., Reidla, M., et al. (2008). Paleo-eskimo mtDNA genome reveals matrilineal discontinuity in greenland. *Science (New York, N.Y.), 320*(5884), 1787-1789.

Gonzalez, A. M., Garcia, O., Larruga, J. M., & Cabrera, V. M. (2006). The mitochondrial lineage U8a reveals a paleolithic settlement in the basque country. *BMC Genomics, 7*, 124.

Gonzalez, A. M., Larruga, J. M., Abu-Amero, K. K., Shi, Y., Pestano, J., & Cabrera, V. M. (2007). Mitochondrial lineage M1 traces an early human backflow to africa. *BMC Genomics, 8*, 223.

Harpending, H. C., Batzer, M. A., Gurven, M., Jorde, L. B., Rogers, A. R., & Sherry, S. T. (1998). Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences of the United States of America, 95*(4), 1961-1967.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America, 106*(23), 9362-9367.

HUGO Pan-Asian SNP Consortium, Abdulla, M. A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S. K., et al. (2009). Mapping human genetic diversity in asia. *Science (New York, N.Y.), 326*(5959), 1541-1545.

Hurles, M. E., Veitia, R., Arroyo, E., Armenteros, M., Bertranpetit, J., Perez-Lezaun, A., et al. (1999). Recent male-mediated gene flow over a linguistic barrier in iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *American Journal of Human Genetics, 65*(5), 1437-1448.

Ingman, M., Kaessmann, H., Paabo, S., & Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature, 408*(6813), 708-713.

International HapMap Consortium. (2003). The international HapMap project. *Nature, 426*(6968), 789-796.

International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449*(7164), 851-861.

Iriondo, M., Barbero, M. C., Izagirre, N., & Manzano, C. (1997). Data on six short-tandem repeat polymorphisms in an autochthonous basque population. *Human Heredity, 47*(3), 131-137.

Iriondo, M., Barbero, M. C., & Manzano, C. (2003). DNA polymorphisms detect ancient barriers to gene flow in basques. *American Journal of Physical Anthropology, 122*(1), 73-84.

Izagirre, N., & de la Rua, C. (1999). An mtDNA analysis in ancient basque populations: Implications for haplogroup V as a marker for a major paleolithic expansion from southwestern europe. *American Journal of Human Genetics, 65*(1), 199-207.

Keinan, A., Mullikin, J. C., Patterson, N., & Reich, D. (2009). Accelerated genetic drift on chromosome X during the human dispersal out of africa. *Nature Genetics, 41*(1), 66-70.

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., et al. (2003). The genetic heritage of the earliest settlers persists both in indian tribal and caste populations. *American Journal of Human Genetics, 72*(2), 313-332.

Krings, M., Salem, A. E., Bauer, K., Geisert, H., Malek, A. K., Chaix, L., et al. (1999). mtDNA analysis of nile river valley populations: A genetic corridor or a barrier to migration? *American Journal of Human Genetics, 64*(4), 1166-1176.

Laayouni, H., Calafell, F., & Bertranpetit, J. (2010). A genome-wide survey does not show the genetic distinctiveness of basques. *Human Genetics, 127*(4), 455-458.

Lambert, C. A., Connelly, C. F., Madeoy, J., Qiu, R., Olson, M. V., & Akey, J. M. (2010). Highly punctuated patterns of population structure on the X chromosome and implications for african evolutionary history. *American Journal of Human Genetics, 86*(1), 34-44.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, N.Y.), 319*(5866), 1100-1104.

Loogvali, E. L., Roostalu, U., Malyarchuk, B. A., Derenko, M. V., Kivisild, T., Metspalu, E., et al. (2004). Disuniting uniformity: A pied cladistic canvas of mtDNA haplogroup H in eurasia. *Molecular Biology and Evolution, 21*(11), 2012-2021.

Maca-Meyer, N., Gonzalez, A. M., Larruga, J. M., Flores, C., & Cabrera, V. M. (2001). Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics, 2*, 13.

Maca-Meyer, N., Gonzalez, A. M., Pestano, J., Flores, C., Larruga, J. M., & Cabrera, V. M. (2003). Mitochondrial DNA transit between west asia and north africa inferred from U6 phylogeography. *BMC Genetics, 4*, 15.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., et al. (2005). Single, rapid coastal settlement of asia revealed by analysis of complete mitochondrial genomes. *Science (New York, N.Y.), 308*(5724), 1034-1036.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., et al. (1999). The emerging tree of west eurasian mtDNAs: A synthesis of control-region sequences and RFLPs. *American Journal of Human Genetics, 64*(1), 232-249.

McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., et al. (2008). Runs of homozygosity in european populations. *American Journal of Human Genetics, 83*(3), 359-372.

Mellars, P. (2006). Going east: New genetic and archaeological perspectives on the modern human colonization of eurasia. *Science (New York, N.Y.), 313*(5788), 796-800.

MOURANT, A. E. (1947). The blood groups of the basques. *Nature, 160*(4067), 505.

Nalls, M. A., Simon-Sanchez, J., Gibbs, J. R., Paisan-Ruiz, C., Bras, J. T., Tanaka, T., et al. (2009). Measures of autozygosity in decline: Globalization, urbanization, and its implications for medical genetics. *PLoS Genetics, 5*(3), e1000415.

Neckelmann, N., Li, K., Wade, R. P., Shuster, R., & Wallace, D. C. (1987). cDNA sequence of a human skeletal muscle ADP/ATP translocator: Lack of a leader peptide, divergence from a fibroblast translocator cDNA, and coevolution with mitochondrial DNA genes. *Proceedings of the National Academy of Sciences of the United States of America, 84*(21), 7580-7584.

Nelis, M., Esko, T., Magi, R., Zimprich, F., Zimprich, A., Toncheva, D., et al. (2009). Genetic structure of europeans: A view from the north-east. *PloS One, 4*(5), e5472.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nature Reviews.Genetics, 8*(11), 857-868.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., et al. (2008). Genes mirror geography within europe. *Nature, 456*(7218), 98-101.

Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., et al. (2006). The mtDNA legacy of the levantine early upper palaeolithic in africa. *Science (New York, N.Y.), 314*(5806), 1767-1770.

Pereira, L., Richards, M., Goios, A., Alonso, A., Albarran, C., Garcia, O., et al. (2005). High-resolution mtDNA evidence for the late-glacial resettlement of europe from an iberian refugium. *Genome Research, 15*(1), 19-24.

Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J., et al. (2003). Joining the pillars of hercules: MtDNA sequences show multidirectional gene flow in the western mediterranean. *Annals of Human Genetics, 67*(Pt 4), 312-328.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38*(8), 904-909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics, 81*(3), 559-575.

Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R., et al. (2004). Where west meets east: The complex mtDNA landscape of the southwest and central asian corridor. *American Journal of Human Genetics, 74*(5), 827-845.

Rando, J. C., Pinto, F., Gonzalez, A. M., Hernandez, M., Larruga, J. M., Cabrera, V. M., et al. (1998). Mitochondrial DNA analysis of northwest african populations reveals genetic exchanges with european, near-eastern, and sub-saharan populations. *Annals of Human Genetics, 62*(Pt 6), 531-550.

Rando, J. C., Pinto, F., Gonzalez, A. M., Hernandez, M., Larruga, J. M., Cabrera, V. M., et al. (1998). Mitochondrial DNA analysis of northwest african populations reveals genetic exchanges with european, near-eastern, and sub-saharan populations. *Annals of Human Genetics, 62*(Pt 6), 531-550.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., et al. (2010). Ancient human genome sequence of an extinct palaeo-eskimo. *Nature, 463*(7282), 757-762.

Reidla, M., Kivisild, T., Metspalu, E., Kaldma, K., Tambets, K., Tolk, H. V., et al. (2003). Origin and diffusion of mtDNA haplogroup X. *American Journal of Human Genetics, 73*(5), 1178-1190.

Repping, S., Skaletsky, H., Brown, L., van Daalen, S. K., Korver, C. M., Pyntikova, T., et al. (2003). Polymorphism for a 1.6-mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nature Genetics, 35*(3), 247-251.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., et al. (2000). Tracing european founder lineages in the near eastern mtDNA pool. *American Journal of Human Genetics, 67*(5), 1251-1276.

Rodriguez-Ezpeleta, N., Alvarez-Busto, J., Imaz, L., Regueiro, M., Azcarate, M. N., Bilbao, R., et al. (2010). High-density SNP genotyping detects homogeneity of spanish and french basques, and confirms their genomic distinctiveness from other european populations. *Human Genetics, 128*(1), 113-117.

Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics, 73*(6), 1402-1422.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., et al. (2002). Genetic structure of human populations. *Science (New York, N.Y.), 298*(5602), 2381-2385.

Rosser, Z. H., Zerjal, T., Hurles, M. E., Adojaan, M., Alavantic, D., Amorim, A., et al. (2000). Y-chromosomal diversity in europe is clinal and influenced primarily by geography, rather than by language. *American Journal of Human Genetics, 67*(6), 1526-1543.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature, 449*(7164), 913-918.

Salas, A., Comas, D., Lareu, M. V., Bertranpetit, J., & Carracedo, A. (1998). mtDNA analysis of the galician population: A genetic edge of european variation. *European Journal of Human Genetics : EJHG, 6*(4), 365-375.

Schaffner, S. F. (2004). The X chromosome in population genetics. *Nature Reviews.Genetics, 5*(1), 43-51.

Semino, O., Magri, C., Benuzzi, G., Lin, A. A., Al-Zahery, N., Battaglia, V., et al. (2004). Origin, diffusion, and differentiation of Y-chromosome haplogroups E

and J: Inferences on the neolithization of europe and later migratory events in the mediterranean area. *American Journal of Human Genetics, 74*(5), 1023-1034.

Soares, P., Achilli, A., Semino, O., Davies, W., Macaulay, V., Bandelt, H. J., et al. (2010). The archaeogenetics of europe. *Current Biology : CB, 20*(4), R174-83.

Spencer, C. C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., et al. (2006). The influence of recombination on human genetic diversity. *PLoS Genetics, 2*(9), e148.

Stevanovitch, A., Gilles, A., Bouzaid, E., Kefi, R., Paris, F., Gayraud, R. P., et al. (2004). Mitochondrial DNA sequence diversity in a sedentary population from egypt. *Annals of Human Genetics, 68*(Pt 1), 23-39.

Stevanovitch, A., Gilles, A., Bouzaid, E., Kefi, R., Paris, F., Gayraud, R. P., et al. (2004). Mitochondrial DNA sequence diversity in a sedentary population from egypt. *Annals of Human Genetics, 68*(Pt 1), 23-39.

Teo, Y. Y., Sim, X., Ong, R. T., Tan, A. K., Chen, J., Tantoso, E., et al. (2009). Singapore genome variation project: A haplotype map of three southeast asian populations. *Genome Research, 19*(11), 2154-2162.

Tian, C., Plenge, R. M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., et al. (2008). Analysis and application of european genetic substructure using 300 K SNP information. *PLoS Genetics, 4*(1), e4.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., et al. (2009). The genetic structure and history of africans and african americans. *Science (New York, N.Y.), 324*(5930), 1035-1044.

Torroni, A., Achilli, A., Macaulay, V., Richards, M., & Bandelt, H. J. (2006). Harvesting the fruit of the human mtDNA tree. *Trends in Genetics : TIG, 22*(6), 339-345.

Torroni, A., Richards, M., Macaulay, V., Forster, P., Villems, R., Norby, S., et al. (2000). mtDNA haplogroups and frequency patterns in europe. *American Journal of Human Genetics, 66*(3), 1173-1177.

Umetsu, K., & Yuasa, I. (2005). Recent progress in mitochondrial DNA analysis. *Legal Medicine (Tokyo, Japan), 7*(4), 259-262.

Underhill, P. A., & Kivisild, T. (2007). Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics, 41*, 539-564.

Vicoso, B., & Charlesworth, B. (2006). Evolution on the X chromosome: Unusual patterns and processes. *Nature Reviews.Genetics, 7*(8), 645-653.

Wallace, D. C. (1995). 1994 william allan award address. mitochondrial DNA variation in human evolution, degenerative disease, and aging. *American Journal of Human Genetics, 57*(2), 201-223.

Wallace, D. C., Ye, J. H., Neckelmann, S. N., Singh, G., Webster, K. A., & Greenberg, B. D. (1987). Sequence analysis of cDNAs for the human and bovine ATP synthase beta subunit: Mitochondrial DNA genes sustain seventeen times more mutations. *Current Genetics, 12*(2), 81-90.

Walter, R. C., Buffler, R. T., Bruggemann, J. H., Guillaume, M. M., Berhe, S. M., Negassi, B., et al. (2000). Early human occupation of the red sea coast of eritrea during the last interglacial. *Nature, 405*(6782), 65-69.

White, T. D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G. D., Suwa, G., et al. (2003). Pleistocene homo sapiens from middle awash, ethiopia. *Nature, 423*(6941), 742-747.

Zlojutro, M., Roy, R., Palikij, J., & Crawford, M. H. (2006). Autosomal STR variation in a basque population: Vizcaya province. *Human Biology; an International Record of Research, 78*(5), 599-618.