Study of human genetic diversity

Inferences on population origin and history

Marc Haber

DOCTORAL THESIS UPF / 2013

THESIS DIRECTOR

Dr. David Comas

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES

Universitat Pompeu Fabra Barcelona

To my mother Leila,
to Mira,
and to Rana.

Acknowledgements

In preparation of this thesis, I am mostly grateful to David Comas and Pierre Zalloua. David for being a mentor and a friend, for giving me the opportunity to work in his team, and for finding a solution to every challenge we have faced in the past few years. Pierre's contribution to this work has been paramount; in guidance, financial support, and data sharing. The data for the first two studies presented here was generated in Pierre's Beirut lab.

I am thankful to David Soria for the long constructive discussions on the genetics of populations studied here. David had important contributions to the Levant study (section 3.2).

I would like to thank Sonia Youhanna for her substantial help on the Levant study. Sonia has recruited 2,000 subjects that we have used in this work and also contributed to the planning of the study.

I am also thankful to Kamal Jaroudy for his insightful comments on the history of the Levant and the different groups inhabiting this region.

I would like to thank Karima Fadhlaoui-Zid for her insights on the genetics of North African populations and for recruiting Libyan samples used in the study (section 3.3).

I thank Maziar Ashrafian Bonab for recruiting the Afghan samples (section 3.1) and for his insights into the demographics of Central

Asia populations. Also on this subject, I would like to thank Dan Platt for his help and guidance which improved the Afghan study notably.

Thanks to Michella Ghassibe-Sabbagh and Angelique Salloum for collecting and managing the Type 2 diabetes samples used in section 4.3.

Thanks to Francesc Calafell, Jaume Bertranpetit, and Chris Tyler-Smith for giving guidance and suggestions throughout the development of this work.

I would like to thank Txema Heredia and Jordi Rambla for their help and support on using the Cadaques cluster at the IBE. Thanks to Judit Sainz for the administrative support and for making things run smooth in the lab.

This work has been substantially improved through continuous discussions and debates with other lab members: Begoña Martínez-Cruz, Isabel Mendizabal, Laura Botigué, Arturo Silveyra, Michael Ducore, and Paula Sanz. I am thankful, you have made this an enjoyable experience.

Abstract

Patterns of human genetic diversity suggest that all modern humans originated from a small population in Africa that expanded rapidly 50,000 years ago to occupy the whole world. While moving into new environments, genetic drift and natural selection affected populations differently, creating genetic structure. By understanding the genetic structure of human populations, we can reconstruct human history and understand the genetic basis of diseases. The work presented here contributes to the ongoing effort to catalogue human genetic diversity by exploring populations that have been underrepresented in genetic studies. We use variations on the genomes of populations from Central Asia, the Near East, and North Africa to reconstruct the history of these populations. We find that climate change and geography appear to be major factors shaping genetic diversity. In addition, we identify recent cultural developments and historical events that have influenced admixture and gene flow between populations, leading to the genetic diversity observed in humans today.

Resum

Els patrons de diversitat genètica humana suggereixen que els humans van sorgir d'un petit grup a l'Àfrica que es va expandir ràpidament fa uns 50,000 anys per tot el planeta. En migrar cap a nous hàbitats, la deriva genètica i la selecció natural van afectar de manera diferencial les poblacions, generant una estructura genètica. Mitjançant la comprensió de l'estructura genètica de les poblacions podem reconstruir la història humana i entendre la base genètica de les malalties. Aquest treball contribueix a l'esforç continu de catalogar la diversitat genètica humana explorant poblacions poc representades en altres estudis genètics. Hem utilitzat variacions al llarg del genoma de poblacions d'Àsia Central, Orient Mitjà i el Nord d'Àfrica per tal de reconstruir la seva història. Hem observat que canvis climàtics i geogràfics semblen ser els factors principals que han modelat la diversitat genètica. A més, hem identificat esdeveniments culturals i històrics recents que afavorit les barreges i el flux genètic entre poblacions, generant la diversitat genètica observada avui en dia.

Preface

Traditionally, information about the origin of human populations came from fossil and archeological data. But over the last half century, genetics has played an increasingly important role in our understanding of human evolution. Advances in genetics have provided new tools for approving or rejecting hypotheses about our past. Today, one can expect new discoveries in human history and evolution to emerge from a genetics lab as from archeological sites. Indeed, fascinating new discoveries about our origins appeared recently through studying human genomes which allowed resolving questions about human evolution previously deemed inscrutable. One thrilling example, which substantially revised previous models on our evolution, is the finding that our ancestors have admixed with archaic humans which appear to share more genetic variants with present-day non-African humans than with present-day humans in sub-Saharan Africa. Such findings were made possible through rapid advancement of sequence and genotyping technologies which facilitated efficient survey of archaic and modern human populations. The flow of data allowed the construction of more accurate mathematical models that can infer past processes from modern populations diversity. The implications are vast, not only to understand how we came to be the way we are but also to speculate about the future evolution of our species.

The current work aspires to contribute to the ongoing effort to catalogue human genetic diversity. We study new populations in a region that extends from Central Asia to North Africa where major developments have marked the history of modern humans. We infer

on the history of these populations using uniparental and genome-wide markers, providing new demographic insights essential to the understanding of human evolution in this region.

Table of contents

List of figures

1. Introduction

The introduction starts with a brief review section on the biological and demographic factors that shape genetic diversity. The second section discusses the use of genetic variation to infer on the origin and history of populations. The third section presents our current understanding of the origin and dispersal of modern humans. Also in this section there is an introduction on the demographic history of populations studied in this thesis.

1.1 Processes shaping genetic diversity

Population genetics studies changes in allele frequencies in a population by considering the mechanisms of evolutionary processes over time and space. Investigating and understanding these processes allow us to construct models that interpret present day population diversity and consequently deduce past population processes such as growth rate, divergence, and admixture. Genetic variation is generated continuously by the mutational process. Variation is then governed by factors such as gene flow, genetic drift, and natural selection. These processes will be summarized in this section.

1.1.1 Mutation

Mutation is the source of all genetic variation since it is the only process generating new alleles. It can occur somatically in diseases such as cancer, however these somatic cases are not inherited and are not of evolutionary interest. Evolutionary significant

polymorphisms occur in the transmitted germ line from one individual to another. Single nucleotide polymorphisms (SNPs) arise by insertion, deletion or point mutation involving only one nucleotide. Analysis of trios estimates that the per generation base pair mutation rate is 1-1.2 × 10^{-8} (The 1000 Genomes Project Consortium, 2010). Mutations occur in the paternal line substantially higher than in the maternal line (ratio = 3.9) (Kong *et al.*, 2012). The diversity in mutation rate of SNPs is dominated by the age of the father at conception of the child. The effect is an increase of about two mutations per year. An exponential model estimates paternal mutations doubling every 16.5 years (Kong *et al.*, 2012) (Figure 1).

Because mutation rate is low, it is unlikely that the same mutation at a given site will occur more than once. Therefore copies of the same variant are often identical by descent from a common ancestor in which the mutation first arose. This is known as the infinite allele model where the possibility of back mutations and recurrent mutations is ignored (Kimura & Crow, 1964). For slow mutating SNPs, the infinite allele model appears to be a close approximation of biological reality. However, microsatellites (short segments of DNA tandemly repeated several times, STRs) are much more polymorphic than SNPs. There may be mutational variation within copies of the repeated units, and the copy number changes rapidly. The high mutation rate of the STRs, which introduce fast variation into a population, made them a useful tool for the elucidation of human population history and for forensic purposes. For STRs, a stepwise mutation model (Kimura & Ohta, 1978) enables deriving a

formula for the equilibrium distribution of allelic frequencies in a finite population when selectively neutral alleles are produced in stepwise fashion.

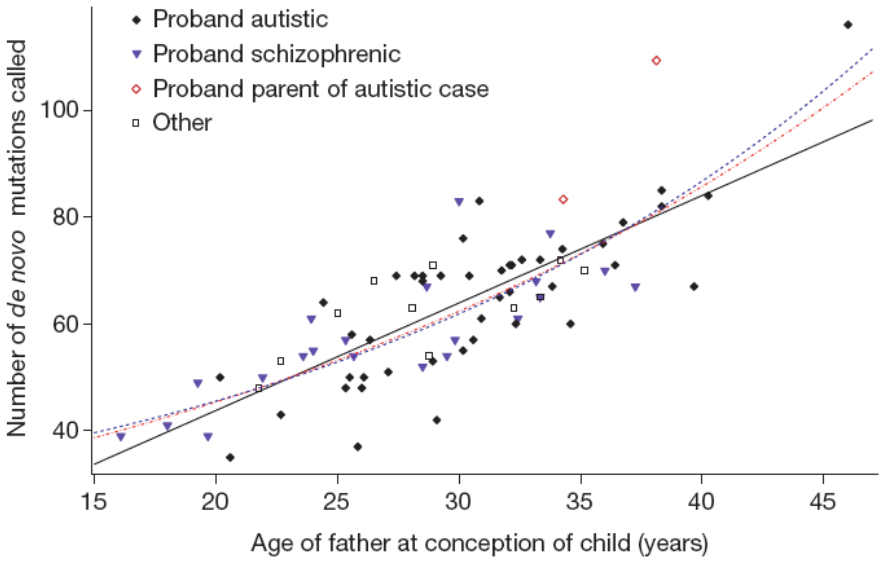

Figure 1. Father's age and number of de novo mutations. The number of de novo mutations called is plotted against father's age at conception of child for 78 trios. The solid black line denotes the linear fit. The dashed red curve is based on an exponential model fitted to the combined mutation counts. The dashed blue curve corresponds to a model in which maternal mutations are assumed to have a constant rate of 14.2 and paternal mutations are assumed to increase exponentially with father's age. From (Kong *et al.*, 2012)

1.1.2 Recombination

Genetic recombination produces new combination of alleles when segments of DNA are exchanged between homologous chromosomes during meiosis. Variations accumulate over time along a chromosome in a stepped hierarchical way, producing new combinations of alleles at neighboring loci known as haplotypes.

Recombination's shuffling process continually generates novel combinations of alleles, allowing an almost infinite number of possible haplotypes. The interaction between loci in a haplotype is often expressed in terms of linkage disequilibrium (LD). The loci are in LD if their respective alleles do not associate randomly. In the absence of evolutionary forces, other than random mating and Mendelian segregation, recombination will act over successive generations to reduce the amount of LD between two physically linked markers. LD decays exponentially along the time axis at a rate that depends on the linkage distance or recombination fraction. The recent flow of genetic data have allowed fine-scale estimation of recombination rates which revealed hotspots of recombination interspersed with stretches of relatively little recombination. The consequence is a block-like structure with blocks showing high internal LD separated by other blocks by low LD between them (International HapMap Consortium, 2005).

Evolutionary forces that shape variation, such as selection, admixture, and drift, can also influence LD. Consequently, population history can generate marked differences in the extent of LD between populations (Reich *et al.*, 2001) (Figure 2).

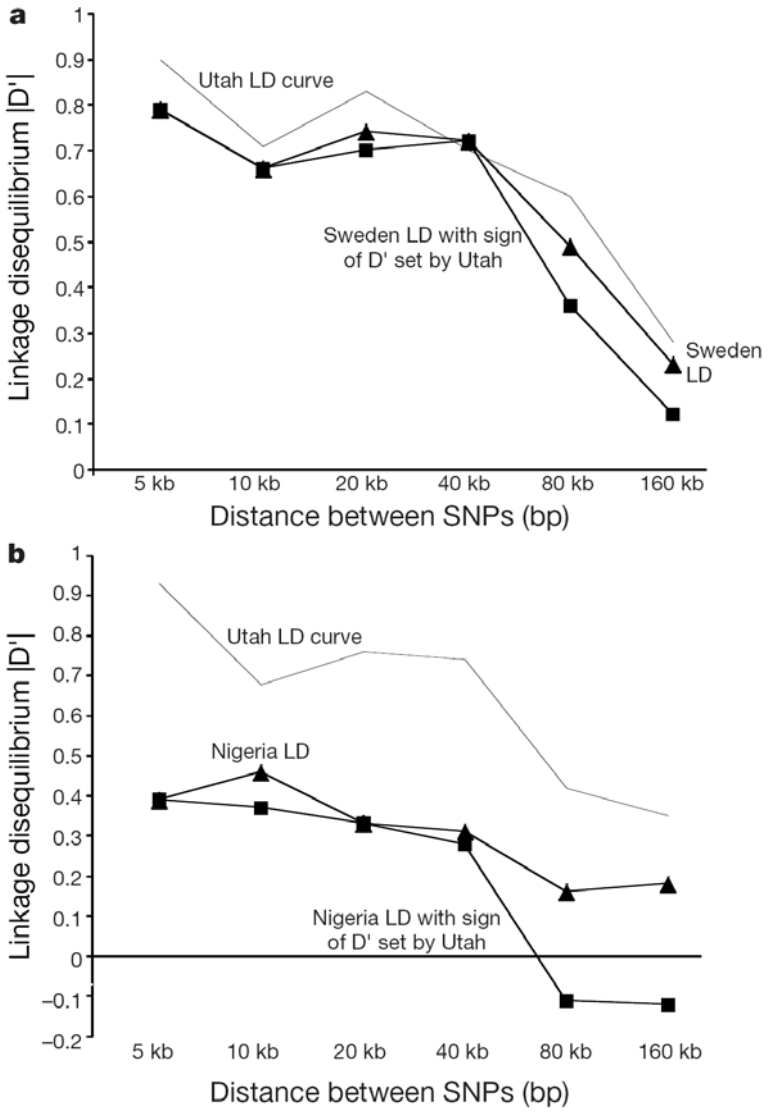

Figure 2. LD curve for Swedish and Yoruban samples. a) In Sweden, average D' is nearly identical to the average |D'| values up to 40-kb distances, and the overall curve has a similar shape to that of the Utah population (thin line in a and b). b) LD extends less far in the Yoruban sample, with most of the long-range LD coming from a single region, HCF2. From (Reich *et al.*, 2001)

1.1.3 Migration and gene flow

The movement of genetic variation over space is known as gene flow. It is the consequence of migration from one inhabited area to another and the reproductive success of the migrant in his/her new location. There are several models of migration: The *n*-island model (Wright, 1931) assumes a metapopulation is split into equal sized islands with equal migration rates (Figure 3). The stepping-stone model (Kimura & Weiss, 1964) introduces the idea of geographical substructure by only allowing gene flow between adjacent subpopulations (Figure 3). Finally, the isolation by distance model (Wright, 1943) considers a continuous population were mating choices are limited by distance, leading to genetic similarity in neighborhoods. Nevertheless, migration processes are far more complex than these models allow. Migrants are rarely a random sample of their source population; they are often age-structured, sexbiased, and related to one another (kin-structured migration) (Jobling *et al.*, 2013)

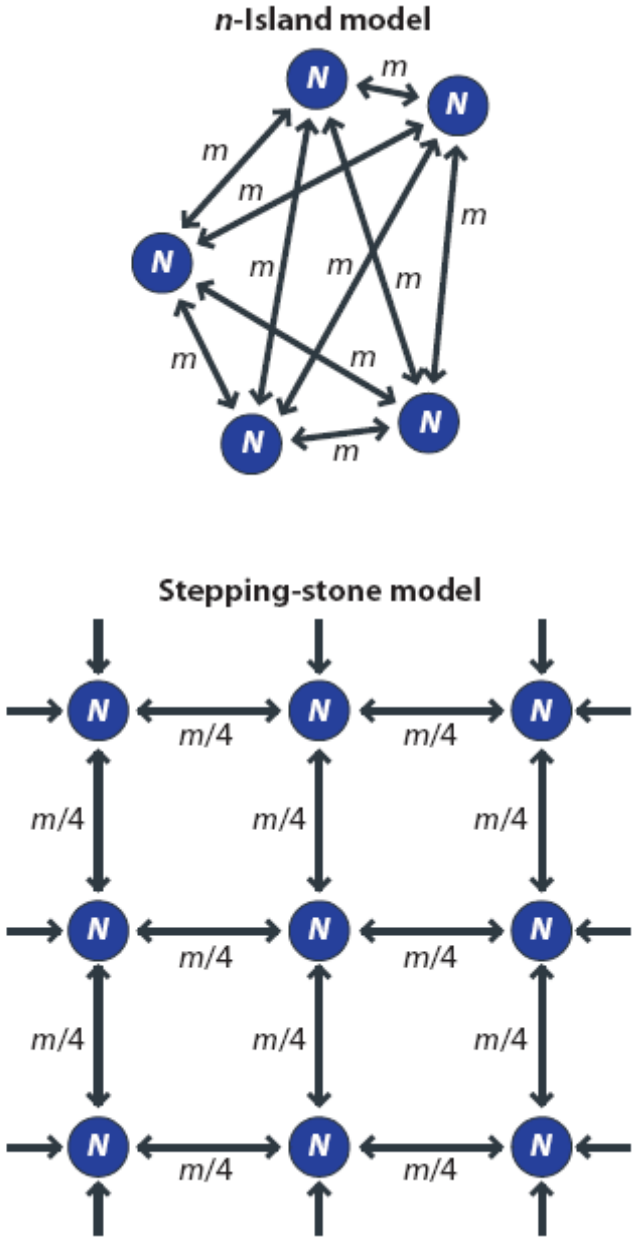

Figure 3. The n-island and steppingstone models of gene flow. Each diagram represents one of a family of models: the N-island model, and the two-dimensional stepping-stone model, also known for obvious reasons as the lattice model. N, population size; m, rate of exchange of genes per generation. From (Jobling *et al.*, 2013)

1.1.4 Genetic drift

Genetic drift refers to the random change in allele frequencies from one generation to the next (Wright, 1931). This occurs by random chance alone since individuals in a population have different reproductive contributions to the next generation. Genetic drift is calibrated by several factors such as population size and growth. It eventually leads to the fixation or loss of specific variants. This occurs faster in a small population accompanied by severe reduction in overall variability in this population (Figure 4). Genetic drift has its greatest effect in population isolates, such as religious sects (like the Druze in the Levant) and island populations.

A basic model that describes genetic drift is known as the Wright-Fisher model (Fisher, 1930, Wright, 1931). It assumes that populations are of constant size, mating is random and generations do not overlap.

The size of an idealized Wright-Fisher population is represented by Wright's concept of effective population size (*N*e) which experiences the same amount of genetic drift as the population under study. A simple model of drift can be constructed by considering a population of diploid individuals (2*Ne*) where each gene copy produces an infinite number of gametes. The probability that two gametes chosen randomly are identical by descent represents the homozygosity of a population (*F*) and is equal to 1/(2*Ne*). *F* then increases over time as

$$F_{t+1} = \frac{1}{2Ne} + \left(1 - \frac{1}{2Ne}\right) F_t$$

The probability that two gametes are not identical by descent represent heterozygosity *H* and calculated as *H*=1-*F* meaning that *H* will decrease over time. The consequence is that, in finite populations, variation is eliminated by drifting to reach fixation and leaving a single allele (*H*=0), unless diversity is maintained by mutation, gene flow or natural selection.

Populations that experience demographic events such as isolation, bottlenecks, and founder effects are subject to strong drift processes, leading to reduced genetic diversity.

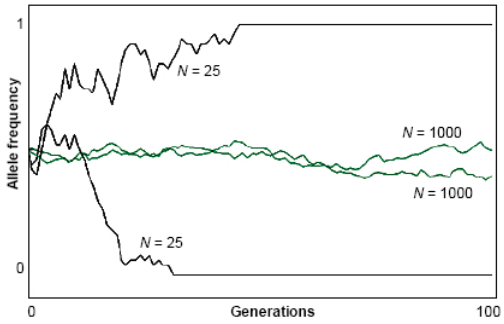

Figure 4. Genetic drift in populations of different sizes. Simulations show the allele frequency from generation to generation of populations of size either 25 or 1000. Each starting from the same initial allele frequency (0.5). The allele rapidly becomes either fixed or lost in small populations, whereas more subtle variations are seen in the larger populations.

1.1.5 Natural selection

Natural selection is defined as the differential reproduction of genetically distinct individuals within a population. It is governed by differences among individuals in traits such as mortality, fertility, mating success and the viability of the offspring, which are collectively referred to as components of fitness. The fitness of a

genotype is a measure of the individual's ability to survive and reproduce. The evolutionary success of an individual is determined by its relative fitness to other genotypes in the population.

Non-neutral mutations can reduce the fitness of the carrier and will be selected against (purifying selection), or can increase fitness giving a selective advantage to its carriers and undergo positive, or diversifying selection.

The fitness in diploid organisms is determined by the interaction between the two alleles at a locus. Several models can be considered including overdominance and underdominance where the heterozygote has the highest or the lowest fitness respectively. Overdominance creates a balanced polymorphism increasing variability in a population. One classic example of balanced polymorphisms in humans is a mutation in the haemoglobin gene which confers protection against malaria when heterozygous, but causes sickle-cell diseases when homozygous (Rosenthal, 2011).

1.1.6 The neutral theory

The neutral theory states that the great majority of evolutionary changes at the molecular level are caused by random drift of selectively neutral mutations (Figure 5). It was first proposed by Motoo Kimura in 1968 based on two observations: a high rate of amino acid substitutions and high amount of polymorphism in diverse organisms (Kimura, 1968). The theory met at first strong opposition from orthodox Darwinians. However, with the flow of sequence data starting the late 1970s, supporting evidence for the neutral theory increased. It is clear today that the neutral theory

does not entirely explain the observed genetic variation and adaptation in humans and other species, nevertheless it provides the theoretical framework to study variation and evolution and a powerful null model to detect selection.

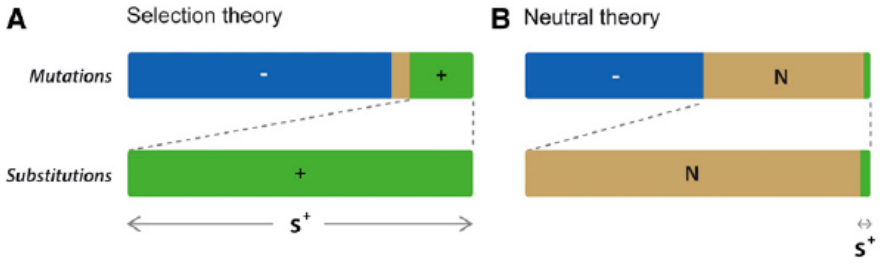

Figure 5. Selection versus Neutral theory. A) Darwin's theory postulated the existence of deleterious (-) and advantageous (+) changes. Deleterious mutations are immediately rejected by negative (or purifying) selection and neutral mutations are neglected. All the substitutions have a positive selection coefficient $s > 0$. B) The neutral theory postulated the existence of an important fraction of neutral mutations (N) and a very small fraction of advantageous mutations. Neutral mutations are fixed by random drift and constitute the majority of substitutions. A very small minority of substitutions have $s > 0$. From (Razeto-Barry *et al.*, 2012)

1.1.7 Nonrandom mating in humans

Selection of mates in humans is influenced by phenotypic traits such as skin color or stature or by abstract traits such as cultural affiliation. The genetic consequence is that the allelic frequencies drift almost independently within the homogeneous subgroups of the population and new mutations stay within the limits of the subgroups.

The most common nonrandom mating pattern among humans is the positive assortative mating in which individuals mate with others who are "like themselves". The net evolutionary effect of this

mating pattern is a progressive decrease in genetic diversity and increase in the number of homozygous genotypes in the population. A less common nonrandom mating pattern is the negative assortative mating in which individuals select mates “different from themselves”. This leads to increased diversity and a progressive increase in the frequency of heterozygous genotypes. However, the genetic consequences of human nonrandom mating patterns are more complex than the above two models. For example, in section 3.2 we show that individuals belonging to the Muslim faith in the Levant region practice positive assortative mating by choosing mates that are also Muslims. However, their mate selection can extends beyond national and regional borders which increased their genetic diversity compared to other groups in the region.

1.2 Making demographic inferences from genetic diversity

Demographic processes such as divergence, migration, and admixture leave an imprint on the genetic diversity of populations. This section summarizes the methods and tools that measure genetic variation and infer on populations' relationships as well as on past processes.

1.2.1 The genetic markers

The genetic markers used to study populations have changed considerably over the last 40 years. The major aim remains to construct a realistic understanding of human population changes in space and time.

Genetic markers showed differences between human populations first by using blood groups and protein types. These "classical markers" preceded DNA based markers and comparisons were based only on allele frequencies since the molecular basis of the polymorphisms was unknown. With the advance of polymerase chain reaction (PCR) methods, molecular markers were introduced and allowed, in addition of allele frequencies measure, an assessment of evolutionary distances between alleles at a locus.

The first molecular markers focused on the mitochondrial DNA (mtDNA) and the nonrecombining region of the Y chromosome (NRY). mtDNA is inherited maternally and transmitted from a mother to her descendents while the NRY is inherited paternally passing down from father to son only. These markers pass intact from generation to generation, escaping recombination and the

chromosome shuffling process. They change only by mutation and therefore present a simple record of their history and allow an easy construction of unique phylogenies.

The development of genome-wide SNP chips, and later advances in whole-genome sequencing, allowed hundreds of thousands of polymorphisms to be surveyed collectively. This permitted the assessment of evolutionary processes without a particular sex bias. In addition, the effective population size of genome-wide markers is expected to be four times that of mtDNA and NRY which pushes back the time of coalescence, giving more in-depth view of the population history. Also, this makes the genome-wide markers less prone to genetic drift and when analyzed collectively the markers should present a selectively neutral overall picture.

The studies presented in this thesis investigated human populations by using NRY and genome-wide markers. NRY markers consisted of SNPs (Y-SNPs) which designate haplogroups or clades that are informative on populations ancestries and allow comparisons to a well resolved phylogeography. In addition, NRY markers consisted of microsatellite (Y-STRs) forming haplotypes that evolve at high rate and allow within haplogroup diversity comparisons.

The genome-wide markers were obtained from SNP arrays (Illumina) consisting of over 550,000 markers, the majority of which are tag SNPs derived from release 20 of the HapMap Project. On average, there is 1 SNP every 5.5-6.2 kb (depending on the population) across the genome.

1.2.2 Measures of molecular diversity

A simple measure of diversity is by counting the number of an allele at a locus or by calculating the mean number of alleles over a range of loci. This summary statistics allow comparisons between populations and can infer on some demographic processes such population bottlenecks or admixture events when the mean number of alleles is low or high.

Another measure of diversity is gene diversity statistics (Nei, 1973), which is a measure of the proportion of polymorphic loci across the genome. The unbiased estimator for gene diversity (Nei & Roychoudhury, 1974) is given as:

$$\mathrm{H} = \frac{\mathrm{n}}{\mathrm{n}\text{-}1}\left(1 - \sum_{i=1}^{I} p_i^2\right)$$

Where n is the number of gene copies and p_i is the frequency of the *i*th allele.

A similar measurement to Nei's statistics to measure nucleotide diversity is represented by π and describes heterozygosity at a nucleotide position:

$$\pi = \frac{\mathrm{n}}{\mathrm{n}-1}\left(\sum x_i x_j \delta_{ij}\right)$$

which represents the probability that two copies of the same nucleotide (x) chosen randomly from a set of sequences (n) will be different from one another.

Assuming neutrality (mutation-drift equilibrium) and infinite sites model, π would be equal to $\theta = 4N_e\mu$ for diploid, $\theta = 2N_e\mu$ for haploid.

Other measures of diversity take into account haplotypes or average comparisons between pairs of individuals/molecules.

Early comparisons of human populations using simple measurements of genetic diversity were able to present plausible models on human origins and migrations. For example, (Vigilant *et al.*, 1991) used the average number of nucleotide differences per 100 bp for comparing pairs of individuals from a population. Using just 610 nucleotides and 189 individuals, they showed that Africans had the greatest genetic diversity and suggested an African origin of human mtDNA evolution.

1.2.3 Measures of apportionment of diversity

The division of a meta-population into partially isolated non-randomly mating subpopulations results in differential fixation and lost of alleles in these populations. Subpopulation divergence results in a deficiency of heterozygotes in the meta-population. This process creates a hierarchical population structure in which the apportionment of diversity among these different subpopulations can be measured.

A common measure of population structure is F_{st} derived from the fixation indices proposed by Wright (Wright, 1951) to measure deviation of observed heterozygote frequencies from those expected under Hardy-Weinberg equilibrium. F_{st} can be defined as:

$F_{st} = (H_T - H_S)/H_T$

where H_T and H_S are the expected heterozygosity in the meta and subpopulation respectively.

F_{st} varies between 0 and 1. When the subpopulations are genetically close, as in continuous admixture, high gene flow, or recent split, F_{st} will be closer to 0. When subpopulations are highly differentiated F_{st} will be close to 1.

A measure analogous to Wright's F_{st} is R_{st} which is inferred from microsatellite data (Slatkin, 1995). It differs in taking a generalized stepwise mutation model of the mutation process at microsatellite loci.

A widely used measure of the extent of population subdivision is the Analysis of Molecular Variance (AMOVA) (Excoffier *et al.*, 1992). This method produces estimates of variance components and F-statistic analogs (Φ-statistics), reflecting the correlation of haplotypic diversity at different levels of hierarchical subdivision. AMOVA can be applied to any data where genetic distances between alleles can be calculated.

1.2.4 Measures of genetic distances between populations

There are a number of measures for genetic distances, these usually make assumptions on the mechanisms driving population divergence.

A common measure for genetic distances is a modification of the fixation index summarized in the previous section. As a genetic distance between two populations F_{st} is defined as:

$F_{st} = V_p/p(1\text{-}p)$

where p and V_p are the mean and variance of gene frequencies between the two populations respectively.

Another widely used measure of genetic distance is Nei's standard genetic distance (Nei, 1972). It assumes that differences arise due to mutation and genetic drift and is defined as:

$$D = -\ln\left(\sum x_i y_i / \sum x_i^2 \left(y_i^2\right)^{1/2}\right)$$

where x_i and y_i are the frequency of ith allele drawn from two populations.

Cavalli-Sforza and Edwards defined a measure that assumes genetic drift only and populations are considered as points in a mutli-dimensional space (Cavalli-Sforza & Edwards, 1967). The measure is suitable for constructing phylogentic trees and has been used frequently with microsatellite data. The distance between two populations is given as the length of the chord joining them:

$$\left(2\sqrt{2/\pi}\right)\sqrt{1-\cos\theta}$$

where

$$\cos\theta = \sum_{i=1}^{k} \sqrt{x_i y_i}$$

where x_i and y_i are the frequency of ith allele drawn from two populations and $2/\pi$ is a scaling constant.

1.2.5 Clustering methods for inference on population structure and admixture

Grouping individuals or populations based on genetic resemblance has the potential for detecting population structure, shared ancestry, and admixture events. There are generally two types of clustering methods that can be used:

Distance based methods

These methods are based on calculating pairwise genetic distances between a set of individuals or populations. The information from the resulting distance matrix is then reduced to two or three dimensions by a method such as Multidimensional Scaling (MDS) and represented graphically for identification of clusters. Principal Component Analysis (PCA) is another method that uses the raw data from allele frequencies rather than genetic distances. Each extracted component/axis represents a percentage of the total variation between individuals or populations (Figure 6). Another method is by constructing a tree using, for example, the Unweighted Pair-Group Method with Arithmetic mean (UPGMA) or a Neighbor-Joining (NJ) method (Figure 6) that use an iterative clustering process to combine individuals or populations that have the least genetic distance. PHYLIP (Felsenstein, 1989) is a widely used package for construction of such trees.

Model-based methods

These methods assume that observations from each cluster are random draws from some parametric model. It attempts to assign

individuals to *K* ancestral populations on the basis of their genotypes, while simultaneously estimating population allele frequencies using maximum-likelihood or Bayesian statistical methods (Pritchard *et al.*, 2000). These methods are implemented in softwares such as STRUCTURE (Pritchard *et al.*, 2000) and ADMIXTURE (Figure 6) (Alexander *et al.*, 2009).

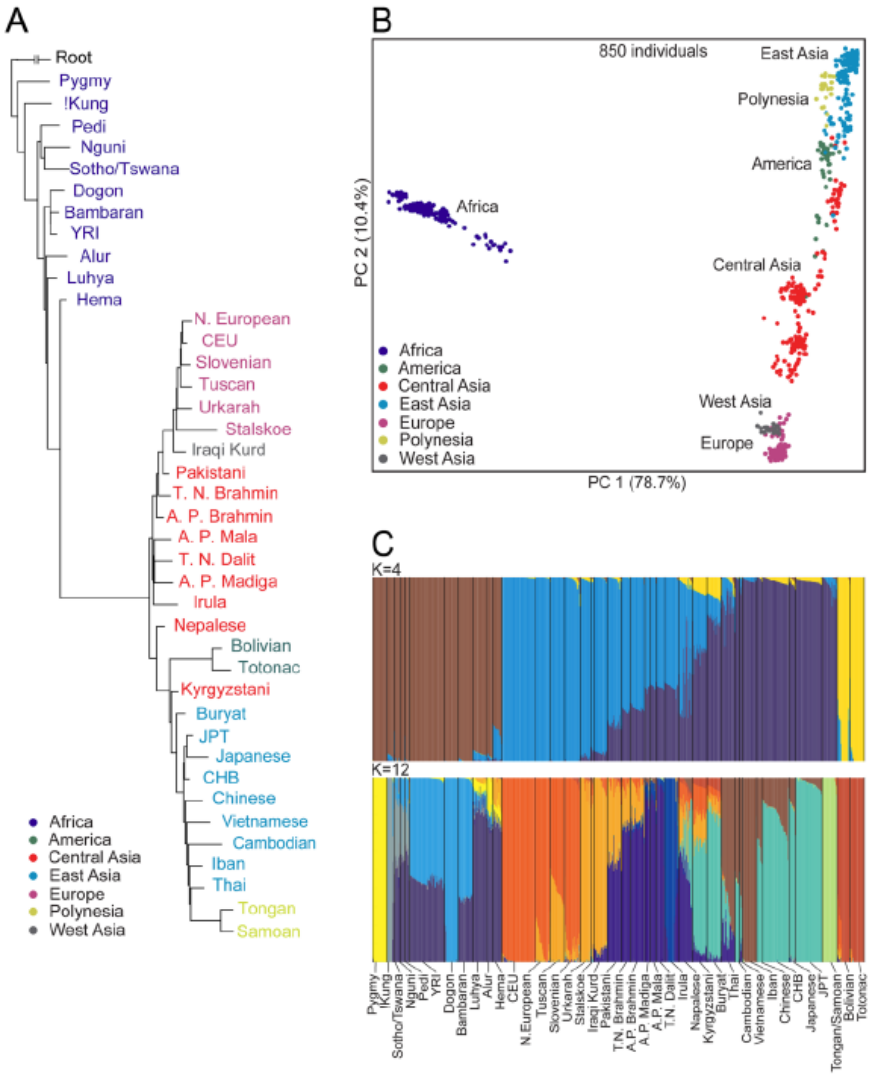

Figure 6. Examples of clustering methods. A) Neighbor-joining tree of 40 Populations color-coded based on their continental origins. B) Principal component analysis showing the first two principal components. Each individual is represented by one dot color-labeled according to regional origin. C) Individual grouping inferred by ADMIXTURE. Results from K = 4 and K = 12 are shown. Each individual's genome is represented by a vertical bar composed of colored sections, where each section represents the proportion of an individual's ancestry derived from one of the K ancestral populations. From (Xing *et al.*, 2010)

1.2.6 Dating population divergence, admixture, and time to the most recent common ancestor

Inferences from genetic distances

A useful measure of time of population divergence using the fixation index can be obtained by making the assumption that divergence is due to drift alone. The impact of genetic drift is largely determined by the effective population size (N_e) and time (t):

$$t = -2Ne\ln(1 - F_{st})$$

Similarly, Nei's genetic distance is also related linearly to time:

$D = 2\alpha t$

Where α is the rate of fixation of nucleotide substitution.

Inferences from allelic diversity of nonrecombining lineages

Rho (ρ) and the Averaged Squared Distance (ASD) dating can be used to estimate the time to the most common ancestor (TMRCA) using Y chromosome and mtDNA haplotypes. ρ requires the construction of a rooted phylogeny (often a median network) using intra-allelic diversity. The ρ statistics represents the average number of mutational changes between root and sample counted from the network. ρ is related to time by:

$\rho = \mu t$

ASD is calculated from the variance in the data without the need to construct a phylogeny. It assumes a stepwise model and therefore is appropriate to microsatellite data only.

Inferences from Linkage Disequilibrium

The time of divergence can be estimated from observed linkage disequilibrium (LD) decay between genetic markers in different populations. The correlation of LD between populations will be reduced each generation by an amount that depends on the recombination rate between markers (Sved *et al.*, 2008). Divergence time (t) can be is estimated by:

$t = -[\ln(\text{mean } r_1\, r_2) - \ln(\text{mean adjusted } r^2)]/(2c)$

Where r_1 and r_2 are the LD correlations in the two populations respectively, r^2 is the squared correlation adjusted for sample size, and c is the known recombination fraction.

The rate of exponential decline of admixture LD can be used to estimate the date since admixture between two populations. The method implemented in a software called *Rolloff* (Moorjani *et al.*, 2011) uses statistic for LD between a pair of markers and a weight that reflects their allele frequency differentiation in the ancestral populations. An estimate of the date is obtained by examining the correlation between pairs of markers as they become separated by increasing genetic distances.

Inferences from model-based approaches

To capture full information from the genetic data, models should incorporate the effect of processes such as genetic drift, mutation, and recombination. These models frequently apply a Markov chain Monte Carlo (MCMC)-based Bayesian inference which allows estimation of population sizes and population divergence times. The

estimates are based on the state of the posterior distribution after iteratively sampling probability distributions. The models require priors such as the effective population size and the mutation rate.

One implementation using nonrecombining microsatellite data or mtDNA data is BATWING (Wilson *et al.*, 2003) which generates posterior distributions of genealogical trees incorporating demographic factors such as population growth and historical splitting into isolated sub-populations (Figure 7).

Another implementation, BEAST (Drummond *et al.*, 2012), uses molecular sequences from a population to reconstruct phylogenies and test evolutionary hypotheses including estimating TMRCA of populations.

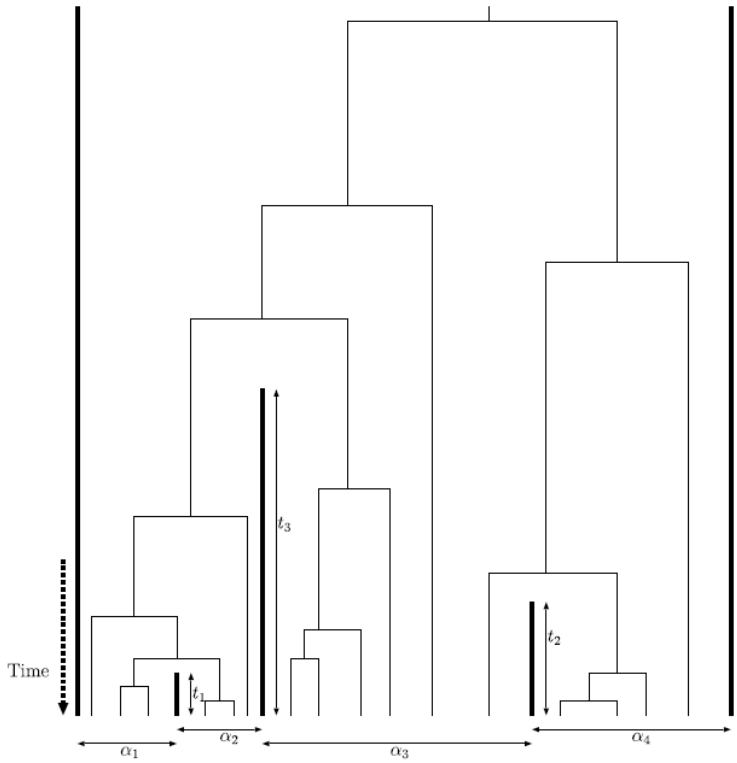

Figure 7. Genealogy under BATWING's 'splitting' model of population subdivision. The single ancestral subpopulation split *t*3 (≡ *ta*) coalescent time units ago into two subpopulations, each of which subsequently split to result in four current subpopulations, whose sizes form proportions α_i , i = 1,...,4, of the total population size *N*; the subpopulation sample sizes are 3, 3, 6 and 4, corresponding to the number of leaves (terminal nodes) (the broken arrow in the left-hand margin indicates the direction of true time; coalescent time runs in the reverse direction). From (Wilson *et al.*, 2003)

1.3 Genetic diversity reveals modern humans' origin and history

Understanding the processes shaping genetic diversity and its appropriate measures summarized in the previous sections, allow translation of genetic information into insights on populations' demographic history. The following section will review what is known today on the origin and expansion of modern humans. In addition, there will be an overview on the demographic history of populations analyzed in this thesis: Afghans, Levantines, and North Africans.

1.3.1 Origin and expansion of modern humans

Theories about the origin of modern humans can be summarized in two main models; multiregional evolution and recent replacement. According to the multiregional model of human evolution, our ancestors encompass the entire Old World population of archaic humans, they evolved during the past 1.5-2 million years in different areas of the Old World, with enough gene flow between these regions to prevent speciation (Thorne & Wolpoff, 1992).

The second model, the replacement model of human evolution states that all modern humans have a relatively recent African origin, with a subsequent dispersal throughout the Old World that completely replaced the existing archaic population (Disotell, 1999). Over the last two decades, archeological and genetic data provided evidence in favor of a recent African origin of modern humans with young time estimates to African most recent common

ancestors. Specifically, 200,000 to 50,000 years ago (ya) fossil record from Africa show a gradual accumulation of anatomically modern osteological features, with the Omo and Herto skulls from Ethiopia, representing the probable immediate ancestors of anatomically modern humans 195,000 to 160,000 ya (Weaver, 2012, McDougall *et al.*, 2005, White *et al.*, 2003). Fossil record also indicate that the first out-of-Africa migration of anatomically modern humans occurred around 100,000 ya (Schwarcz & Grun, 1992). Evidence points specifically to the Levant where modern humans lived around 90,000 ya. It is believed that the range of anatomically modern humans later retracted back to Africa, due to their inability to thrive in the harsher, colder, and more arid non-tropical environment of Late Pleistocene Levant. This is indicated by the lack of archaeological records from the Levant between 90,000 and 50,000 ya (Bar-Yosef, 1992). Furthermore, archaeology indicates these humans appear to have been behaviorally pre-modern and to have been subsequently replaced by Neanderthals.

After the retraction of the range of anatomically modern humans following their first expansion, archaeological and genetic data suggest a date of around 50,000 ya for a second expansion of fully modern humans and the origin of all current non-African populations (Stringer, 2012, Henn *et al.*, 2012b, Quintana-Murci *et al.*, 1999). The routes by which fully modern humans spread out of Africa remain controversial (Balter, 2011, Macaulay *et al.*, 2005), with new archeological and genetic evidence favoring a route over a land bridge crossing what today is the Bab el Mandeb strait separating Djibouti from the Arabian Peninsula at the southern end

of the Red Sea (Fernandes *et al.*, 2012, Mele *et al.*, 2012, Armitage *et al.*, 2011). The expansion was accompanied by a continuous decrease of genetic diversity with increasing distance from Africa and a pattern consistent with a serial founder model where new populations are formed from a subset of the expanding wave outward from Africa. (Mele *et al.*, 2012, Henn *et al.*, 2012b, DeGiorgio *et al.*, 2009, Li *et al.*, 2008). The loss of genetic variation in humans is remarkable, for example, the recent sequencing of 79 great ape genomes identified more than double the number of SNPs from more than a thousand diverse humans (Prado-Martinez *et al.*, 2013, The 1000 Genomes Project Consortium, 2010).

However, the Out of Africa model of human expansion has not been without a twist. Analyses of ancient DNA from two archaic human groups, the Neanderthals and Denisovans, suggest that limited gene flow from these archaic *Homo* species to modern humans occurred in two brief episodes (Green *et al.*, 2010, Reich *et al.*, 2010). One episode occurred at an early stage of the out-of-Africa expansion of modern humans, and the second occurred only in the ancestors of Melanesian populations in Oceania (Figure 8).

These findings show how complexly fascinating is the origin and history of modern humans. Future studies including the one presented in this thesis will eventually lead to a better understanding of the differences that exist between the human groups. This should eventually help in determining how it was that modern humans came to expand dramatically in population size and cultural complexity over the last 60,000 years.

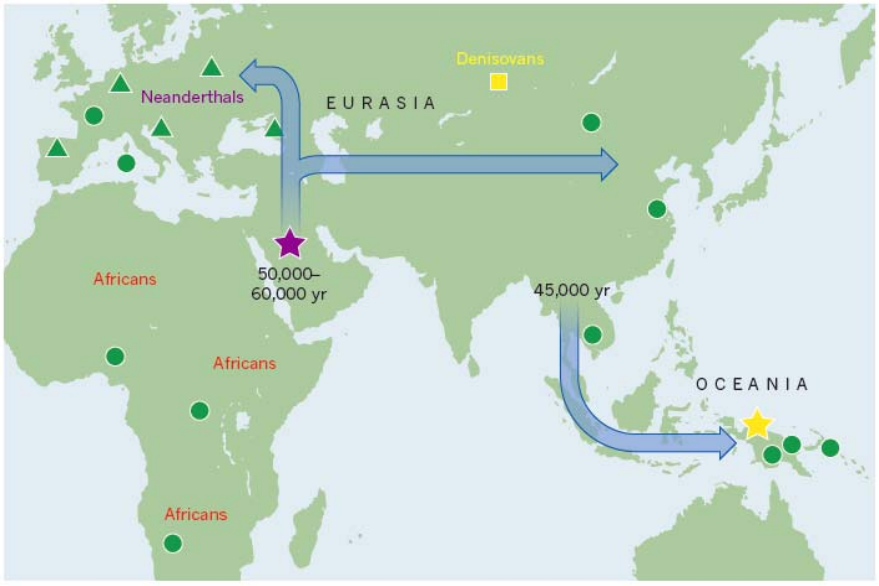

Figure 8. Hypotheses on modern human history. Triangles and circles respectively represent sampling locations of Neanderthal remains and of present-day human genomes. The blue arrows indicate generally accepted major migrations of anatomically modern humans, following their departure from Africa 50,000-60,000 years ago. At this time, there were two primary archaic species in Eurasia, Neanderthals and *Homo erectus*; Reich, Pääbo and co-workers suggest that a third group was also present, represented by the ancient Denisovan genome. From ancient DNA they identify additional putative events involving two episodes of limited gene flow: first, genetic admixture from Neanderthals to modern humans, shortly after the exit from Africa; second, subsequent admixture with the archaic population exemplified by the nuclear DNA extracted from the Denisova finger bone. This second event seems to affect only the ancestors of present-day Melanesians, who are thought to have colonized Papua New Guinea some 45,000 years ago. African populations, both past and present, are genetically highly diverse, as indicated by the multiple labels. From (Bustamante & Henn, 2010)

1.3.2 Demographic history of Afghanistan

Afghanistan is a landlocked country at the intersection of Central Asia, South Asia, and the Middle East and has been a crossroad of human migrations since the Paleolithic era. Due to political instability in the region, archeological excavations in Afghanistan are limited and most findings are reported from works in the 1950's and 1960's. These findings suggest modern humans inhabited northern Afghanistan as early as 50,000 ya. The same area also appear to have been a place for the development of the first agricultural communities (Dupree, 1964), which later probably supported the economy of early urban Bronze Age civilizations in the region (Dupree, 1980). Afghanistan history is marked by the influx of invading armies including Iranians, Greeks, Indians, Mongols, and Arabs.

The present population of Afghanistan contains many diverse elements (Figure 9), the result of large-scale migrations and conquests that influenced its culture and demography. Genetic studies on the Afghan population are very scarce, probably due to the difficulty of sampling in the area and the political sensitivity of the topic. The few existing genetic studies did not attempt to analyze the different Afghan ethnic groups and are limited to either listing of autosomal short tandem repeats (STRs) frequencies (Di Cristofaro *et al.*, 2011, Berti *et al.*, 2005) or Y-chromosome STR analysis in a single ethnic group (Lacau *et al.*, 2011).
The current work attempts for the first time to study the genetic diversity of the different Afghan groups. We consider Afghanistan's

ethnic diversity a unique opportunity to explore how nations and ethnic groups emerge, and how major cultural evolutions and technological developments in human history have influenced modern population structures.

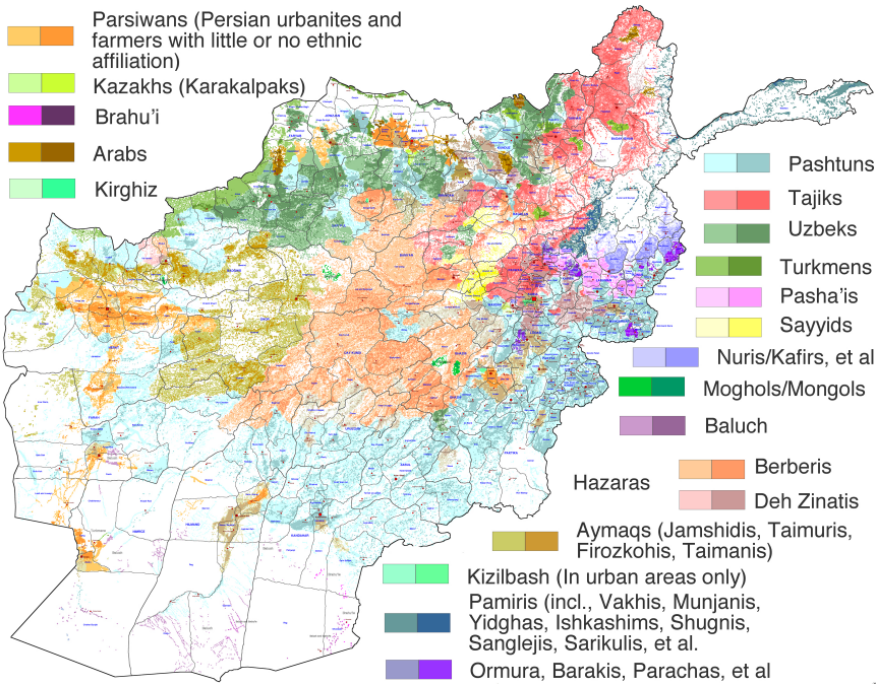

Figure 9. Map of Afghanistan. Colors show geographic distribution of the ethnic groups in Afghanistan. Each group is indicated by a color with two intensities reflecting population density. Modified from gulf2000.columbia.edu. Produced by Dr. Michael Izady.

1.3.3 Demographic history of the Levant

The Levant is a geographical area in the eastern Mediterranean region bounded by Anatolia, Egypt, and the Arabian Desert. Fossil records indicate the Levant has been inhabited since modern humans first ventured out of Africa (Schwarcz & Grun, 1992, Bar-Yosef, 1992). The Levant also sits in a region embracing the earliest agricultural communities, the rise of the first urban cities and the emergence of the first civilizations.

Genetic studies using uniparental markers show that historical events, such as the Islamic expansion and the Crusades, have marked the genomes of modern Levantines (Zalloua *et al.*, 2008). Genome-wide surveys in the Levant are limited to studies assessing the relationship of Diaspora Jewish groups to a Levantine/Middle Eastern origin (Atzmon *et al.*, 2010, Behar *et al.*, 2010). However, the genetic diversity of other groups in the region, their relation to each others and to their neighbors, is still unknown. Indeed, the Levant today is a mosaic of religious groups inhabiting a relatively narrow geography but practicing strong endogamy (Figure 10). Conveniently, the history of the Levant region and its religious groups is meticulously documented. For example, we know that Christianity in most Levant dates back to the first century CE, whereas Islam was brought to the region through the Islamic expansions in 635 CE. In 986 CE, the Druze faith developed as a movement within Islam, and from 1030 CE, a person could only be Druze if born Druze. Therefore, the Levant provides an opportunity to calibrate historical information and genetic data to study how

isolation, admixture, and migration (driven by cultural affiliation) impact the genetic structure of modern populations.

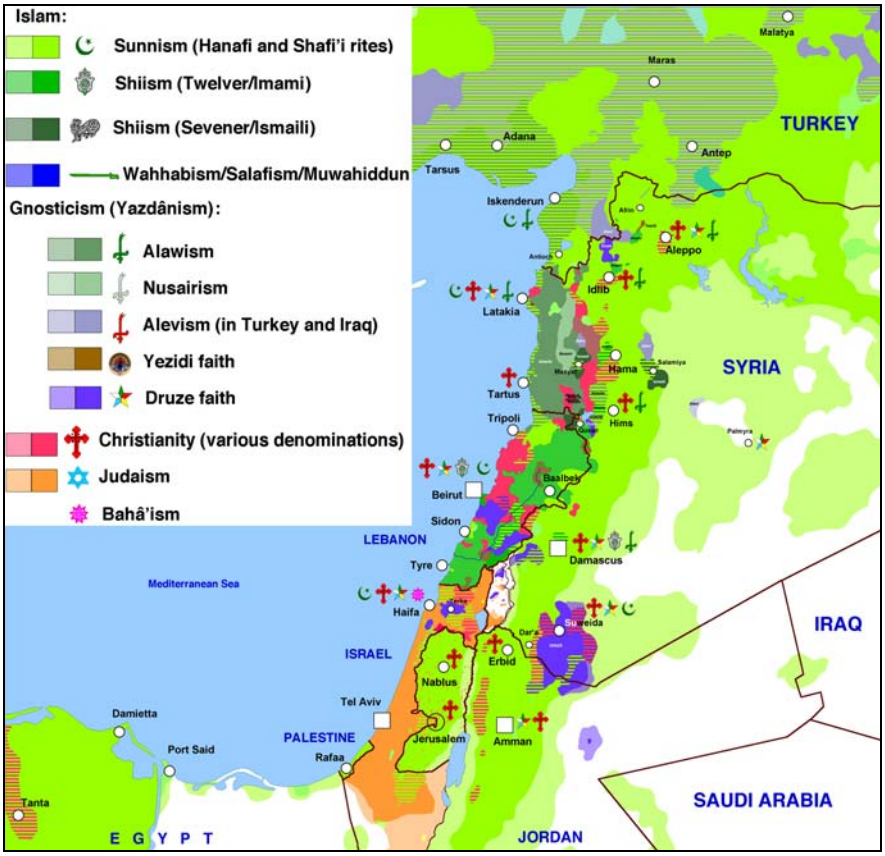

Figure 10. The religious composition of the Levant. Map shows the present distribution of the religious groups in the Levant. Each group is indicated by a color with two intensities reflecting population density. Mixed areas are hachured. Modified from gulf2000.columbia.edu. Produced by Dr. Michael Izady.

1.3.4 Demographic history of North Africa

Most archeologists and geneticists today suggest that Africa is the place of origin of all modern humans. However, other details about the exact location and timing of the emergence of *Homo sapiens* and the routes from which they have expanded out of Africa are still a subject of debate.

The finding of fossilized hominid crania from Ethiopia that are morphologically and chronologically intermediate between archaic African fossils and later anatomically modern Late Pleistocene humans constituted strong evidence of modern human emergence in East Africa (Stringer, 2003, White *et al.*, 2003). However, later archeological findings in North Africa, including that of a 160,000 ya juvenile *Homo sapiens* displaying an equivalent degree of tooth development to modern European children at the same age, highlights the importance of North Africa in the emergence and spread of modern humans (Smith *et al.*, 2007, Balter, 2011). As a result of these findings, a thorough description of North Africa's genetic diversity becomes important to understand human evolution.

The work presented in this thesis takes advantage of the wealth of data and information that have accumulated from past genetic studies on North Africa. It attempts to analyze markers with different inheritance patterns (uniparental and genome-wide) and examines similarities/contrasts in the results, consequently providing a comprehensive description of the evolutionary history of North Africa populations.

2. Objectives

The thesis explores the genetic diversity of modern populations to learn about past demographic processes and consequently reconstruct the history and origin of these populations. Each study had a set of specific aims:

Study of the Afghan population

1- Analyze Y-chromosomal variation in the major ethnic groups of Afghanistan and provide for the first time deep phylogenetic information on Afghan haplogroup memberships.

2- Explore whether the ethnic groups in Afghanistan reflect different social systems that arose in a common population or whether cultural differences are founded on already existing genetic differences.

3- Identify traces of historical movements that influenced the different ethnic groups.

4- Explore how the establishment of the first civilizations in the region affected the present Afghan genetic diversity.

Study of the Levantine population

1- Assess the genome-wide genetic relationships of the Levantines and resolve previous uncertainties about population structure in the Levant region.

2- Determine the factors driving population structure in the Levant by studying the role of geography and religion.

3- Explore how cultural affiliation and transition, through promoting or obstructing admixture between the diverse cultural groups, have impacted the genetic structure of modern populations.

4- Explore culturally and genetically isolated populations in the Levant, like the Christians and Druze, to infer on the past genetic diversity of the Levant region and accurately construct the genetic relationships with neighboring populations.

Study of North African populations

1- Analyze Y-chromosome and genome-wide markers and compare to previous findings from mtDNA in North African populations.

2- Compare North Africans to neighboring populations and study admixture patterns with these populations.

3- Study populations' emergence and split and analyze these demographic processes in light of environmental and historical events.

3. Results

3.1 Afghanistan's Ethnic Groups Share a Y-Chromosomal Heritage Structured by Historical Events

Marc Haber, Daniel E Platt, Maziar Ashrafian Bonab, Sonia C Youhanna, David F Soria-Hernanz, Begoña Martínez-Cruz, Bouchra Douaihy, Michella Ghassibe-Sabbagh, Hoshang Rafatpanah, Mohsen Ghanbari, John Whale, Oleg Balanovsky, R Spencer Wells, David Comas, Chris Tyler-Smith, Pierre A Zalloua & The Genographic Consortium

PLoS ONE. 2012; 7(3): e34288. doi:10.1371/journal.pone.0034288

Abstract

Afghanistan has held a strategic position throughout history. It has been inhabited since the Paleolithic and later became a crossroad for expanding civilizations and empires. Afghanistan's location, history, and diverse ethnic groups present a unique opportunity to explore how nations and ethnic groups emerged, and how major cultural evolutions and technological developments in human history have influenced modern population structures. In this study we have analyzed, for the first time, the four major ethnic groups in present-day Afghanistan: Hazara, Pashtun, Tajik, and Uzbek, using 52 binary markers and 19 short tandem repeats on the non-recombinant segment of the Y-chromosome. A total of 204 Afghan samples were investigated along with more than 8,500 samples from surrounding populations important to Afghanistan's history through migrations and conquests, including Iranians, Greeks, Indians, Middle Easterners, East Europeans, and East Asians. Our results suggest that all current Afghans largely share a heritage derived from a common unstructured ancestral population that could have emerged during the Neolithic revolution and the formation of the first farming communities. Our results also indicate that inter-Afghan differentiation started during the Bronze Age, probably driven by the formation of the first civilizations in the region. Later migrations and invasions into the region have been assimilated differentially among the ethnic groups, increasing inter-population genetic differences, and giving the Afghans a unique genetic diversity in Central Asia.

3.2 Genome-Wide Diversity in the Levant Reveals Recent Structuring by Culture

Marc Haber, Dominique Gauguier, Sonia Youhanna, Nick Patterson, Priya Moorjani, Laura R. Botigué, Daniel E. Platt, Elizabeth Matisoo-Smith, David F. Soria-Hernanz, R. Spencer Wells, Jaume Bertranpetit, Chris Tyler-Smith, David Comas*, Pierre A. Zalloua*

PLoS Genet. 2013; 9(2): e1003316. doi:10.1371/journal.pgen.1003316

Abstract

The Levant is a region in the Near East with an impressive record of continuous human existence and major cultural developments since the Paleolithic period. Genetic and archeological studies present solid evidence placing the Middle East and the Arabian Peninsula as the first stepping-stone outside Africa. There is, however, little understanding of demographic changes in the Middle East, particularly the Levant, after the first Out-of-Africa expansion and how the Levantine peoples relate genetically to each other and to their neighbors. In this study we analyze more than 500,000 genome-wide SNPs in 1,341 new samples from the Levant and compare them to samples from 48 populations worldwide. Our results show recent genetic stratifications in the Levant are driven by the religious affiliations of the populations within the region. Cultural changes within the last two millennia appear to have facilitated/maintained admixture between culturally similar populations from the Levant, Arabian Peninsula, and Africa. The same cultural changes seem to have resulted in genetic isolation of other groups by limiting admixture with culturally different neighboring populations. Consequently, Levant populations today fall into two main groups: one sharing more genetic characteristics with modern-day Europeans and Central Asians, and the other with closer genetic affinities to other Middle Easterners and Africans. Finally, we identify a putative Levantine ancestral component that diverged from other Middle Easterners ~23,700-15,500 years ago during the last glacial period, and diverged from Europeans ~15,900-9,100 years ago between the last glacial warming and the start of the Neolithic.

3.3 Genome-Wide and Paternal Diversity Reveal a Recent Origin of Human Populations in North Africa

Karima Fadhlaoui-Zid*, **Marc Haber***, Begoña Martínez-Cruz, Pierre Zalloua, Amel Benammar Elgaaied, David Comas

Accepted PLoS ONE

Genome-Wide and Paternal Diversity Reveal a Recent Origin of Human Populations in North Africa

Karima Fadhlaoui-Zid[1,2,*], Marc Haber[1,3,*], Begoña Martínez-Cruz[1], Pierre Zalloua[3], Amel Benammar Elgaaied[2], David Comas[1]

[1] Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Spain

[2] Laboratoire de Génétique, Immunologie et Pathologies Humaines, Faculté des Sciences de Tunis, Campus Universitaire El Manar II, Université el Manar, 1068 Tunis, Tunisia

[3] The Lebanese American University, Chouran, Beirut, Lebanon

*Both authors contributed equally

ABSTRACT

The geostrategic location of North Africa as a crossroad between three continents and as a stepping-stone outside Africa has evoked anthropological and genetic interest in this region. Numerous studies have described the genetic landscape of the human population in North Africa employing paternal, maternal, and biparental molecular markers. However, information from these markers which have different inheritance patterns has been mostly assessed independently, resulting in an incomplete description of the region. In this study, we analyze uniparental and genome-wide markers examining similarities or contrasts in the results and consequently provide a comprehensive description of the evolutionary history of North Africa populations. Our results show that both males and females in North Africa underwent a similar admixture history with slight differences in the proportions of admixture components. Consequently, genome-wide diversity show similar patterns with admixture tests suggesting North Africans are a mixture of ancestral populations related to current Africans and Eurasians with more affinity towards the out-of-Africa populations than to sub-Saharan Africans. We estimate from the paternal lineages that most North Africans emerged ~15,000 years ago during the last glacial warming and that population splits started after the desiccation of the Sahara. Although most North Africans share a common admixture history, the Tunisian Berbers show long periods of genetic isolation and appear to have diverged from surrounding populations without subsequent mixture. On the other hand, continuous gene flow from the Middle East made Egyptians

genetically closer to Eurasians than to other North Africans. We show that genetic diversity of today's North Africans mostly captures patterns from migrations post Last Glacial Maximum and therefore may be insufficient to inform on the initial population of the region during the Middle Paleolithic period.

INTRODUCTION

The peopling of North Africa is particularly interesting for anthropologists and human population geneticists due to North Africa's strategic location at a crossroad between Europe, the Middle East and the rest of Africa. The area has been characterized by shifting patterns of human settlements with human movements constrained by the Mediterranean Sea and the Sahara Desert, which might have limited migrations into an east-west direction. However, recent studies have suggested that these barriers might have not been totally impermeable to human movements. Diverse migration and admixture processes appear to have played a pivotal role in shaping the peopling of North Africa since the Middle Paleolithic period. Archaeological data suggest that the earliest modern humans arrived to North Africa around 160,000 years ago (ya) (Smith et al., 2007b). Human settlements dated between 145,000 ya and 40,000 ya were associated with the Aterian lithic industry (Barton et al., 2009, Garcea, 2010), which was replaced by the Iberomaurusian culture during the Last Glacial Maximum (Debénath, 2000). During the Holocene, part of North Africa (mainly Eastern Maghreb) was characterized by the Capsian culture, which developed in situ in the Maghreb and experienced a Neolithic transition in their later phase (Camps, 1974, Camps, 1982). During the historical period, North Africa has been settled successively by diverse populations including Phoenicians, Romans, Vandals and Byzantines. By the end of the 7th century A.D, Arab armies from the Arabian Peninsula arrived to North Africa spreading Islam and the Arabic language in the region. Subsequent migrations of Arab populations

followed, in particular the 10th century saw considerable movement of Bedouins to North Africa (Murdock, 1959, Hiernaux, 1975).

Early genetic studies have identified an Upper Paleolithic component in current northern African populations, and suggested that the Neolithic transition occurred through cultural diffusion (Barbujani et al., 1994, Bosch et al., 1997). Studies using autosomal markers such as short tandem repeats (STRs), polymorphic Alu insertions, HLA class II polymorphisms, and GM and KM allotypes have shown close genetic affinity of North Africans to Eurasian populations and found evidence of gene flow from sub-Saharan populations (Chaabani et al., 1984, Loveslati et al., 2001, Fadhlaoui-Zid et al., 2004a, Abdennaji Guenounou et al., 2006, Fadhlaoui-Zid et al., 2010, Bosch et al., 2000, Cherni et al., 2005a, Coudray et al., 2007, Khodjet-El-Khil et al., 2008, Comas et al., 2000, Flores et al., 2000, Gonzalez-Perez et al., 2003, Ennafaa et al., 2006, Frigi et al., 2011). Recent genome-wide analysis of North Africans found substantial shared ancestry with the Middle East, and to a lesser extent sub-Saharan Africa and Europe. An autochthonous Maghrebi ancestry that increases from east to west across northern Africa was also identified. It was suggested that this ancestry likely derive from ‘‘back-to-Africa’’ gene flow more than 12,000 ya (Henn et al., 2012a). In addition, it has been suggested that recent gene flow between the Middle East and North Africa was probably promoted by shared cultures after the Islamic expansion, increasing genetic similarities between North Africans and Middle Easterners (Haber et al., 2013). Interestingly, genome-

wide analysis also shows that increased genetic diversity in Southern Europe, which is higher than in other regions of the continent, is a result of recent gene flow from North Africa (Botigué et al., 2013).

Analysis of uniparental markers have found two Y-chromosome lineages (E1b1b1a-M78 and E1b1b1b-M81) at high frequency in North African populations, although the origin and emergence of these lineages have been controversial, with some studies suggesting a Paleolithic component (Bosch et al., 2001), while other studies pointing to a Neolithic origin (Arredi et al., 2004, Cruciani et al., 2004, Cruciani et al., 2007, Cruciani et al., 2010, Semino et al., 2004). Middle East influence has also been detected mainly through the presence of haplogroup J in some North African groups. In addition, some studies have reported a limited contribution of sub-Saharan paternal lineages to the North African gene pool (Fadhlaoui-Zid et al., 2011b, Ennafaa et al., 2011). Previous analyzes of mtDNA lineages in North African populations suggest significant Eurasian origins (Fadhlaoui-Zid et al., 2004b, Plaza et al., 2003, González et al., 2006)with lineages dating back to Paleolithic times (Fadhlaoui-Zid et al., 2004b) and with recent gene flow from sub-Saharan Africa linked to slave trade (Harich et al., 2010). mtDNA variations showed an East-West cline accompanied by a genetic discontinuity on the Libyan/Egyptian border, suggesting a differential gene flow in the Nile River Valley (Fadhlaoui-Zid et al., 2011a).

In this study, we complement our previous findings on the maternal lineages by analyzing Y-chromosome and genome-wide markers in

North Africans. We analyze Y-chromosome markers in more than 3,000 samples from African and Eurasian populations including 302 new samples from Libya and Morocco. In addition, we explore recently published genome-wide data from North Africa, the Middle East, and Europe using new methodologies for inference of populations' relations. We use this information in addition to previous findings to present for the first time a complete description of the genetic landscape in North Africa.

MATERIALS AND METHODS

Y-chromosome Analysis

Subjects and Comparative Datasets. We have genotyped 302 unrelated males belonging to the general population of Libya (215) and Central Morocco (87). Genealogical information of the donors was recorded for a minimum of two generations to ascertain their paternal ancestry. All samples were procured with informed consent following the ethical guidelines specified by the Institutional Review Board of the Comitè Ètic d'Investigació Clínica-Institut Municipal d'Assistència Sanitària (CEIC-IMAS) in Barcelona, Spain.

For comparative purposes, additional published samples (2,854) from Africa, the Middle East and Europe were included in the analyses (Table S1). The YCC nomenclature (Karafet et al., 2008) was used throughout the manuscript. The Tunisian populations (Fadhlaoui-Zid et al., 2011b) were pooled into one group since Analysis of the Molecular Variance (AMOVA) showed them to be

genetically homogeneous (variation among groups = 0.70%, $p > 0.05$ and 1.50%, $p > 0.05$ for Y-STR and Y-SNP, respectively).

Genotyping. DNA was extracted from blood samples using a standard phenol/chloroform protocol (Gill et al., 1985) and then quantified using the Quantifiler® Human DNA Quantification Kit (Applied Biosystems). Samples were genotyped with a set of fifty-five Y-chromosome SNPs in a hierarchical method using TaqMan® probes (Applied Biosystems). Real-time PCR was performed using a 7900HT Fast Real-Time PCR System (Applied Biosystems) as previously described (Fadhlaoui-Zid et al., 2011b).

Samples were additionally genotyped for seventeen Y-chromosome STRs using the AmplSTR® Yfiler® PCR Amplification Kit (Applied Biosystems) and a 3130xl Genetic Analyzer (Applied Biosystems).

Statistical analyzes. A graphical representation (contour map) of the geographical distribution of Y-chromosome haplogroups frequencies (Table S2) was plotted using Surfer 8.0 (Golden Software Products).

The phylogenetic relationship between haplotypes belonging to E1b1b1b-M81, E1b1b1a E-M78, J1-M267 and J2-M172 haplogroups was inferred through reduced-median networks using Network 4.5.0.1 (Bandelt et al., 1999). Networks were constructed using markers shared across studies: DYS19, DYS389I, DYS389b, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438 and DYS439. Locus DYS389b was calculated by subtracting the DYS389I from DYS389II (co-amplified fragments).

To study the genetic diversity within populations, we calculated haplotype and haplogroup frequencies, haplogroup and haplotype diversity, and mean number of pairwise differences (MPD), using Arlequin 3.5 (Excoffier & Lischer, 2010). Non-metric multidimensional scaling (MDS) was performed in R (R Development Core Team, 2011) using RST distances between populations computed by Arlequin on DYS19, DYS389I, DYS389b, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439. A principal component analysis (PCA) (Jolliffe, 1986) was performed on relative haplogroup frequencies normalized within populations, centered, and without variance normalization. Since haplogroup resolution was not uniform across studies, the haplogroups were reduced to the most informative derived markers shared across studies.

In order to examine the potential signals of population structure in North African populations, a hierarchical analysis of molecular variance (AMOVA) was carried out grouping the populations according to geographical criteria. The main null hypothesis tested by AMOVA was the non-differentiation of Western and Eastern North African populations. Detailed grouping designs are shown in Table S3. AMOVA analyses were performed with Y-STR haplotypes and Y-SNP haplogroups independently using Arlequin 3.5 (Excoffier & Lischer, 2010).

We have used BATWING (Wilson et al., 2003b) to explore demographic factors such as population growth and historical splitting into sub-populations. We considered a model of exponential growth from a constant-size ancestral population. STRs

used to draw the global phylogenetic tree were those used to plot the MDS. Additional four STRs (DYS448, DYS456, DYS458, GATA H4) were added to the comparisons of North Africans. STRs were assigned observed germline mutation rates (Balaresque et al., 2010). All SNPs were included and contributed to resolve the phylogenetic tree; however BATWING does not use this information for posterior estimates. Priors for initial effective population size (11,000) and growth rate (1.01) that cover wide ranges of possible values were used as previously described (Weale et al., 2002, Rebala et al., 2012) along with a male generation interval of 31 years (Fenner, 2005). A total of 1.5 million Markov chain Monte Carlo (MCMC) samples were kept for inferences of demographic factors. A consensus tree was generated using the Fitch program from the PHYLIP package (Felsenstein, 1989b).

Genome-wide Analysis

Comparative datasets. Samples from North Africa (Henn et al., 2012a), the Middle East (Haber et al., 2013), Europe (Henn et al., 2012a), and Sub-Saharan Africa (Li et al., 2008) were merged. PLINK (Purcell et al., 2007) was used for data management and quality control. Genotyping success rate was set to 99%, sex-linked and mitochondrial SNPs removed, keeping 44,000 SNPs.

Population structure. PCA was performed using smartpca, part of the EIGENSOFT 3.0 package (Patterson et al., 2006). A maximum likelihood tree of human populations with mixture events was plotted using TreeMix (Pickrell & Pritchard, 2012). TreeMix was also used for inference of population admixture implementing three-

population tests (Reich et al., 2009). The PCA and tree were visualized using R (R Development Core Team, 2011).

RESULTS

Paternal lineage composition in North African populations

The paternal lineage distribution in North African populations was compared to neighboring European and Levantine groups (Figure 1A) using 302 new North African samples from Libya and Morocco (Figure S1). As previously reported (Bosch et al., 2001, Arredi et al., 2004, Cruciani et al., 2004, Fadhlaoui-Zid et al., 2011b), the two specific North African haplogroups, E1b1b1a-M78 and E1b1b1b-M81, are predominant in North African populations. The second most frequent haplogroup was J, which has been postulated to have a Middle Eastern origin (Semino et al., 2004). Both J sub-branches, J-M267 and J-M172, were observed in North Africans. Contour maps of haplogroup frequencies show that haplogroup E-M81 is frequent in Northwest Africa but declines towards Egypt and the Levant (Figure 1B). On the other hand, E-M78 and E-M123 are frequent in the Levant and Egypt and decline towards Northwest Africa (Figure 1C and D, respectively). The Middle Eastern haplogroups J-M267 and J-M172 were observed in all samples, although with different distributions. J-M267 (Figure 1E) is prevalent in all North African and Levantine groups, whereas J-M172 is primarily distributed in the Levant and sporadically detected in North Africa and Iberia (Figure 1F).

We have studied the main haplogroups further by constructing reduced-median networks from haplotypes found in each

population. The E-M81 network (Figure S2A) is characterized by a star-like shape centered on the most frequent haplotype that is present in all North African and European populations analyzed. Around 11% of the lineages clustered in specific clades within the network pointing to a high level of diversity throughout the region. The overall haplotype diversity (HD) and mean pairwise difference (MPD) values within haplogroup E-M81 are 0.8398 ± 0.0162 and 2.1693 ± 1.2055, respectively.

E-M78 network (Figure S2B) reveals high diversity within the haplogroup. This clade is mostly found in Middle Eastern populations and Northeastern Africans (27% in Libya and 33% in Egypt). Diversity values within haplogroup E-M78 are higher than for E-M81 (0.9903 ± 0.0017 and 4.1361 ± 2.0666, for HD and MPD respectively).

Network analysis of the J-M267 included 448 haplotypes, mostly from Middle Eastern populations (Figure S2D). J-M267 was found in all North Africans except the Tuareg. All North Africans also shared the modal haplotype with the Levantines. Diversity estimates within haplogroup J-M267 were 0.9524 ± 0.0067 and 2.9387 ± 1.5428 for HD and MPD, respectively.

Haplogroup J-M172 was frequent in Middle Eastern groups (73.9%), and less in Europeans (18.5%) and North Africans (7%) (Figure S2C). J-M172 network shows that clusters are shared mostly between Middle Easterners and Europeans and that most North African lineages stem out from Middle Eastern clusters.

North African paternal population structure

Comparison of the studied populations was first carried out using principal component analysis (PCA) on haplogroup frequencies shown in supplementary Table S2. The first two components account for 55.35% of the variation and reveal a strong geographical clustering of the populations analyzed (Figure 2A). The first component separates sub-Saharan Africans which have higher frequencies of B-M60 A-M91, E-M2, and E*-M96 haplogroups. The first component also shows clustering of the Europeans characterized by R*-M207 and I-M170 and Middle Easterners which have higher frequencies of E-M78, E-M123, J-M267, and J-M172. The second component separates all North African populations except Egyptians from all other populations and shows that E-M81 plays a major role in this structure. The Tuareg appear to be drawn towards sub-Saharans while Egyptians clustered with Middle Easterners close to Palestinians

Genetic affinity between the studied groups was further investigated by calculating pairwise genetic distances (RST) using Y-STR haplotypes. The MDS (Figure 2B) shows a geographical clustering similar to the PCA. The first dimension splits the sub-Saharan Africans from all other populations. The North Africans cluster close to Middle Easterners with Tuareg drawn towards sub-Saharans and Egypt close to Palestinians.

We have further investigated the genetic structure found in North Africa by implementing AMOVA on different geographical clusters (Table S3). A significant genetic heterogeneity was found when all populations were considered as a single group (15.17% for

haplogroups and 11.15% for haplotypes). For comparisons with the mtDNA results from Fadhlaoui-Zid et al (Fadhlaoui-Zid et al., 2011a), two groups were considered in each analysis taking into consideration current geopolitical boundaries. Results show significant variance among groups when Morocco, Algeria and Tunisia were pooled in one group and Libya, Tuareg, Egypt and the Middle East pooled in the second group. Variance among groups decreases but remains significant when Libyans and Tuareg are added to the first group. Conversely, significant differences between groups are lost when Egyptians are added to the North African group (Table S3). This result is also reflected in the PCA and MDS and shows Egypt's strong affinity to the Middle East rather than to North Africa.

To examine population relations and the time depth in which the North African structures have emerged, we employed BATWING to create hypotheses on historical population splitting and coalescent events. BATWING results show that North Africans form their own branch, which is close to Middle Easterners (Figure 3). Egypt appears on the Middle East branch rather than with other North Africans, again in agreement with previous analyses. Our results show that most North Africans emerged around 15,000 ya during the post Last Glacial Maxima warming period (Table S4). Tunisians (Chenini-Douiret Berbers) show older dates and appear to have Paleolithic common ancestors with other North Africans. Population structure within North Africa starts with the splitting of Egypt around 2,800 ya. Tuareg split next from North Africans

around 1,900 ya, followed by the remaining North Africans splitting around 1,000-1,300 ya.

North African genome-wide population structure

PCA on genome-wide SNPs (Figure 4A) shows that North Africans are diverse and closer to Middle Easterners and Europeans than to Sub-Saharan Africans. Egyptians appear the closest to Middle Easterners and Europeans while South Moroccans are drawn towards Sub-Saharans. Tunisians samples (Chenini-Douiret Berbers) form an orthogonal cluster close but distinct from other North Africans which mostly appear in overlapping clusters.

We constructed trees that infer population relationships using TreeMix (Pickrell & Pritchard, 2012). This method estimates both population splits and the possibility of population mixture. First, we build a maximum-likelihood tree setting the position of the root at the Yoruba (Figure 4B). South Moroccans and Saharawi appear close to Yoruba while Egyptians are on a branch leading to Middle Easterners and Basque. Next, we set TreeMix to allow migration edges (m) and test by increasing m sequentially up to m=20. The initial tree structure remains mostly unchanged when migration edges are added. All North Africans except Tunisians appear admixed from an ancestral population to Yoruba. For figure clarity, we show plot m=6 and the migration edges weights (Figure S3A). When m>6 the tree shows admixture among North Africans as well admixture with Middle Easterners/Europeans. To visually identify aspects of ancestry not captured by the tree at m=6, we plot the residuals of the model's fit (Figure S3B). Positive residuals indicate

populations where the fit might be improved by adding additional edges. TreeMix results show that relatedness of the tested populations cannot be explained by a simple tree; therefore we apply a 3-population test to all populations to measure treeness in the previous results. A negative value from $f3$(A;B,C) implies that population A derives from at least two different groups that are related to B and C. Table S5 shows the two lowest values for each North African population. All North Africans except Tunisians appear to be a mixture of populations related to Yoruba and Eurasians (Basque and Lebanese Christians). Tunisians, Yoruba, Basque, and Lebanese Christians appear to be related to other groups by a simple tree implying a history of divergence without subsequent mixture.

DISCUSSION

The anthropological interest in North Africa as a crossroad between three continents and as a stepping-stone outside Africa has led to numerous studies describing the genetic landscape of the human population in this region. These studies used paternal, maternal, and biparental molecular markers to investigate population structure in North Africa. However, information from these markers which have different inheritance patterns has been mostly assessed independently, resulting in an incomplete description of North Africa populations. In this study, we analyze uniparental and genome-wide markers proved informative for inferring population origin and history. We explore our populations by examining similarities or contrasts in the results from these markers and

consequently provide a thorough description of the evolutionary history of North Africa populations.

Our results from the maternally inherited mtDNA genome (Fadhlaoui-Zid et al., 2011a) and the paternally inherited Y-chromosome show that both males and females in North Africa underwent a similar admixture history and both are today a mixture of African and Eurasian lineages with more affinity towards the out-of-Africa populations than to sub-Saharan Africans. We should note here that although the pattern of admixture with the surrounding regions is similar in males and females, the demographic processes or historical events driving these admixtures could have been different. Also, differential sexual gene flow might have resulted in differences in the proportions of admixture components resulting in source lineage frequency differences (Fadhlaoui-Zid et al., 2011a). Nevertheless, we show that a generally similar admixture history in male and female phylogenies consequently reflected on the entire genome diversity, resulting in genome-wide SNPs showing comparable patterns to uniparental markers, placing North Africans close to Eurasians. Furthermore, admixture tests using genome-wide SNPs also show that most North Africans are a mixture of populations related to current Africans and Eurasians.

Although recent cultural expansions from the Middle East, like the Islamic expansion, could have introduced new lineages to North Africa and facilitated admixture between populations from both regions, our results show that the North African component mostly formed much earlier. This is shown in the admixture tests where Basque and Lebanese Christians but not Lebanese Muslims formed

potential source populations to North Africans. In particular, Lebanese Christians were shown to have been isolated for at least the last 2,000 years and were proposed to be genetically close to the ancestral population of the Levant region from which current Europeans diverged ~15,900–9,100 ya between the last glacial warming and the start of the Neolithic (Haber et al., 2013). Our coalescence time estimate for the paternal lineages in North Africa is ~15,000 ya for most populations. These dates coincide with major environmental changes in North Africa following the full glacial hyperarid conditions during the Last Glacial Maxima. Humid conditions started in North Africa ~14,500 ya transforming the area into a verdant landscape vegetated with annual grasses and shrubs which attracted hunter-gatherers who spread into the region (Brovkin & Claussen, 2008, Kropelin et al., 2008, Bar-Yosef, 1987). This period was accompanied by cultural connection between the Middle East and North Africa as suggested by the lithic similarity between the regions (Kropelin et al., 2008).

The gradual termination of the African Humid Period started ~6,000 ya establishing today's North Africa desert ecosystem ~2,700 ya(Kropelin et al., 2008). The desiccation of the Sahara accompanied by large-scale dust mobilization from 4,300 ya could have limited population spread and gene flow in the region, hypothetically triggering populations' divergence and structure. Our Bayesian analysis of population splits suggest North African populations started splitting ~2,800 ya (95%CI= 1,300-4,600 ya). Egypt appears to have split first from North Africa with dates coinciding with the kingdom decline in power and conquests by

Assyrians and Persians. Our results from both uniparental and autosomal markers show that today's Egyptians are genetically closer to Eurasians than to other North Africans, probably a consequence of Egypt's and the Middle East's long established interaction through conquests and trades. Tuareg split next from North Africans around 1,900 ya, followed by the remaining North Africans splitting around 1,000-1,300 ya which coincide with the Islamic expansion arriving to North Africa.

Although most North Africans appear as an admixture of populations from the surrounding regions, the Tunisian Berbers show long periods of genetic isolation, allowing a distinctive genetic component to evolve. Unlike other North Africans, our admixture tests propose that Berbers diverged from surrounding populations without subsequent mixture. We show that coalescence time estimate from paternal lineages are pushed back ~15,000 years when Tunisians (Berbers and general population) are included in the analyses suggesting an early upper Paleolithic ancestral population with most North Africans (~30,000-44,000 ya).

There has been recent interest in North Africa as a source for modern human migrations after most early research studying the origins of Homo sapiens focused on the fossils of East Africa. Recent studies of hominin fossils from northwestern Africa present strong evidence of resemblances and possible evolutionary connections with fossils representing migrations out of Africa between 130,000 and 40,000 ya (Balter, 2011b). Our analysis of modern North Africans shows that most populations emerged

recently from admixture of Africans and Eurasians and therefore are ineffective in resolving questions about ancient human expansions. Genetic isolates, like the Tunisian Berbers analyzed here, could provide some insights on early human movements in North Africa. However, information from today's populations is limited by factors such as migration, admixture, drift, and selection pressure. We show that genetic diversity of today's North Africans mostly captures patterns from migrations post Last Glacial Maximum with no traces of genetic continuity with the first human settlers in the region. Therefore, reconstruction of modern humans' history would probably require analysis of indigenous ancient DNA from human fossils.

ACKNOWLEDGEMENTS

We thank Dr. Nejib Naoui for his help with sample collection and all the DNA donors who made this study possible. We also thank Paula Sanz, Mònica Vallés, and the Genomic Core Facility at the UPF for their valuable technical help and advice. This study was supported in parts by Spanish Government MCINN grant CGL2010-14944/BOS and Programa de Cooperación Interuniversitaria e Investigación Científica (AECI), Spanish Ministry of Foreign Affairs and Cooperation grants A75180/06, A/8394/07, B/018514/08, A1/040218/11

REFERENCES

1. Smith TM, Tafforeau P, Reid DJ, Grun R, Eggins S, et al. (2007) From the cover: earliest evidence of modern human life history in North African early Homo sapiens. Proc Natl Acad Sci USA 104: 6128–6133.

2. Barton RNE, Bouzouggar A, Collcutt SN, Schwenninger J-L, Clark-Balzan L (2009) OSL dating of the Aterian levels at Grotte de Dar es-Soltan I (Rabat, Morocco) and possible implications for the dispersal of modern Homo sapiens. Quaternary Sci Rev 28.

3. Garcea EAA (2010) The spread of Aterian peoples in North Africa. In: Garcea, E.A.A., editor. South-Eastern Mediterranean Peoples Between 130,000 and 10,000 years ago. Oxford: Oxbow Books.

4. Debénath A (2000) Le peuplement préhistorique du Maroc: données récentes et problèmes. L'anthropologie 104: 131-145.

5. Camps G (1974) Les civilisations préhistoriques de l'Afrique du Nord et du Sahara. Paris: Doin.

6. Camps G (1982) Beginnings of pastoralism and cultivation in north-west Africa and the Sahara: origins of the Berbers. In: The Cambridge History of Africa Vol1: from the earliest times to c500 BC, JD Clark, ed Cambridge: Cambridge University Press: 548-612.

7. Murdock GP (1959) Africa, Its Peoples and their Culture History. New York, Toronto, London: McGraw-Hill Book Company.

8. Hiernaux J (1975) The people of Africa. New York: Charles Scribner's Sons.

9. Barbujani G, Pilastro A, De Domenico S, Renfrew C (1994) Genetic variation in North Africa and Eurasia: neolithic demic diffusion vs. Paleolithic colonisation. Am J Phys Anthropol 95: 137-154.

10. Bosch E, Calafell F, Perez-Lezaun A, Comas D, Mateu E, et al. (1997) Population history of north Africa: evidence from classical genetic markers. Hum Biol 69: 295-311.

11. Chaabani H, Helal AN, van Loghem E, Langaney A, Benammar Elgaaied A, et al. (1984) Genetic study of Tunisian Berbers. I. Gm, Am and Km immunoglobulin allotypes and ABO blood groups. J Immunogenet 11: 107-113.

12. Loveslati BY, Sanchez-Mazas A, Ennafaa H, Marrakchi R, Dugoujon JM, et al. (2001) A study of Gm allotypes and immunoglobulin heavy gamma IGHG genes in Berbers, Arabs and sub-Saharan Africans from Jerba Island, Tunisia. Eur J Immunogenet 28: 531-538.

13. Fadhlaoui-Zid K, Dugoujon JM, Elgaaied A, Amor MB, Yacoubi B, et al. (2004a) Genetic diversity in Tunisia: a study based on the GM polymorphism of human immunoglobulins. Hum Biol 76: 559-567.

14. Abdennaji Guenounou B, Loueslati BY, Buhler S, Hmida S, Ennafaa H, et al. (2006) HLA class II genetic diversity in southern Tunisia and the Mediterranean area. Int J Immunogenet 33: 93-103.

15. Fadhlaoui-Zid K, Buhler S, Dridi A, Benammar El Gaaied A, Sanchez-Mazas A (2010) Polymorphism of HLA class II genes in Berbers from Southern Tunisia. Tissue Antigens 76: 416-420.

16. Bosch E, Calafell F, Perez-Lezaun A, Clarimon J, Comas D, et al. (2000) Genetic structure of north-west Africa revealed by STR analysis. Eur J Hum Genet 8: 360-366.

17. Cherni L, Loueslati Yaacoubi B, Pereira L, Alves C, Khodjet-El-Khil H, et al. (2005a) Data for 15 autosomal STR markers (Powerplex 16 System) from two Tunisian populations: Kesra (Berber) and Zriba (Arab). Forensic Sci Int 147: 101-106.

18. Coudray C, Calderon R, Guitard E, Ambrosio B, Gonzalez-Martin A, et al. (2007) Allele frequencies of 15 tetrameric short tandem repeats (STRs) in Andalusians from Huelva (Spain). Forensic Sci Int 168: e21-24.

19. Khodjet-El-Khil H, Fadhlaoui-Zid K, Gusmao L, Alves C, Benammar-Elgaaied A, et al. (2008) Substructure of a Tunisian Berber population as inferred from 15 autosomal short tandem repeat loci. Hum Biol 80: 435-448.

20. Comas D, Calafell F, Benchemsi N, Helal A, Lefranc G, et al. (2000) Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. Hum Genet 107: 312-319.

21. Flores C, Maca-Meyer N, Gonzalez AM, Cabrera VM (2000) Northwest African distribution of the CD4/Alu microsatellite haplotypes. Ann Hum Genet 64: 321-327.

22. Gonzalez-Perez E, Via M, Esteban E, Lopez-Alomar A, Mazieres S, et al. (2003) Alu insertions in the Iberian Peninsula and north west Africa--genetic boundaries or melting pot? Coll Antropol 27: 491-500.

23. Ennafaa H, Amor MB, Yacoubi-Loueslati B, Khodjet el-khil H, Gonzalez-Perez E, et al. (2006) Alu polymorphisms in Jerba Island population (Tunisia): comparative study in Arab and Berber groups. Ann Hum Biol 33: 634-640.

24. Frigi S, Ennafaa H, Ben Amor M, Cherni L, Ben Ammar-Elgaaied A (2011) Assessing human genetic diversity in Tunisian Berber populations by Alu insertion polymorphisms. Ann Hum Biol 38: 53-58.

25. Henn BM, Botigue LR, Gravel S, Wang W, Brisbin A, et al. (2012) Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet 8: e1002397.

26. Haber M, Gauguier D, Youhanna S, Patterson N, Moorjani P, et al. (2013) Genome-wide diversity in the levant reveals recent structuring by culture. PLoS Genet 9: e1003316.

27. Botigué LR, Henn BM, Gravel S, Maples BK, Gignoux CR, et al. (2013) Gene flow from North Africa contributes to differential human genetic diversity in Southern Europe. Proceedings of the National Academy of Sciences USA: in press.

28. Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, et al. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68: 1019-1029.

29. Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, et al. (2004) A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. Am J Hum Genet 75: 338-345.

30. Cruciani F, La Fratta R, Santolamazza P, Sellitto D, Pascone R, et al. (2004) Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. Am J Hum Genet 74: 1014-1022.

31. Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, et al. (2007) Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. Mol Biol Evol 24: 1300-1311.

32. Cruciani F, Trombetta B, Sellitto D, Massaia A, Destro-Bisol G, et al. (2010) Human Y chromosome haplogroup R-V88: a paternal genetic record of early mid Holocene trans-Saharan connections and the spread of Chadic languages. Eur J Hum Genet 18: 800-807.

33. Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, 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. Am J Hum Genet 74: 1023-1034.

34. Fadhlaoui-Zid K, Martinez-Cruz B, Khodjet-el-khil H, Mendizabal I, Benammar-Elgaaied A, et al. (2011b) Genetic structure of Tunisian ethnic groups revealed by paternal lineages. Am J Phys Anthropol 146: 271-280.

35. Ennafaa H, Fregel R, Khodjet-El-Khil H, Gonzalez AM, Mahmoudi HA, et al. (2011) Mitochondrial DNA and Y-chromosome microstructure in Tunisia. J Hum Genet 56: 734-741.

36. Fadhlaoui-Zid K, Plaza S, Calafell F, Ben Amor M, Comas D, et al. (2004b) Mitochondrial DNA heterogeneity in Tunisian Berbers. Ann Hum Genet 68: 222-233.

37. Plaza S, Calafell F, Helal A, Bouzerna N, Lefranc G, et al. (2003) Joining the Pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. Ann Hum Genet 67: 312-328.

38. González AM, Cabrera VM, Larruga JM, Tounkara A, Noumsi G, et al. (2006) Mitochondrial DNA variation in Mauritania and Mali and their genetic relationship to other Western Africa populations. Ann Hum Genet 70: 631-657.

39. Harich N, Costa MD, Fernandes V, Kandil M, Pereira JB, et al. (2010) The trans-Saharan slave trade-clues from interpolation analyses and high resolution characterization of mitochondrial DNA lineages. BMC Evol Biol 10: 138-156.

40. Fadhlaoui-Zid K, Rodriguez-Botigue L, Naoui N, Benammar-Elgaaied A, Calafell F, et al. (2011a) Mitochondrial DNA structure in North Africa reveals a genetic discontinuity in the Nile Valley. Am J Phys Anthropol 145: 107-117.

41. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, et al. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18: 830-838.

42. Gill P, Jeffreys AJ, Werrett DJ (1985) Forensic application of DNA 'fingerprints'. Nature 318: 577-579.

43. Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16: 37-48.

44. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10: 564-567.

45. R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing.

46. Jolliffe I (1986) Principal Coponents Analysis. Second Edition New York, NY: Springer.

47. Wilson IJ, Weale ME, Balding DJ (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. Journal of the Royal Statistical Society A 166, part 2.

48. Balaresque P, Bowden GR, Adams SM, Leung HY, King TE, et al. (2010) A predominantly neolithic origin for European paternal lineages. PLoS Biol 8: e1000285.

49. Weale ME, Weiss DA, Jager RF, Bradman N, Thomas MG (2002) Y chromosome evidence for Anglo-Saxon mass migration. Mol Biol Evol 19: 1008-1021.

50. Rebala K, Martinez-Cruz B, Tonjes A, Kovacs P, Stumvoll M, et al. (2012) Contemporary paternal genetic landscape of Polish and German populations: from early medieval Slavic expansion to post-World War II resettlements. Eur J Hum Genet.

51. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128: 415-423.

52. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5.

53. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-1104.

54. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

55. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190.

56. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8: e1002967.

57. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature 461: 489-494.

58. Brovkin V, Claussen M (2008) Comment on "Climate-driven ecosystem succession in the Sahara: the past 6000 years". Science 322: 1326; author reply 1326.

59. Kropelin S, Verschuren D, Lezine AM, Eggermont H, Cocquyt C, et al. (2008) Climate-driven ecosystem succession in the Sahara: the past 6000 years. Science 320: 765-768.

60. Bar-Yosef O (1987) Pleistocene Connexions between Africa and Southwest Asia: An Archaeological Perspective. The African Archaeological Review 5: 29-38.

61. Balter M (2011) Was North Africa The Launch Pad For Modern Human Migrations? Science 331: 20-23.

FIGURE LEGENDS

Figure 1. Frequency of the major Y-chromosome haplogroups in North Africa and surrounding regions. Intensity of the colors reflects the frequency of a haplogroup in the studied populations. A) Location of the analyzed populations. B-F) Frequency distribution of haplogroups E-M81, E-M78, E-M123, J-M267, and J-M172 respectively.

Figure 2. Y-chromosome population structure. A) Principal component analysis of haplogroups frequencies. B) Multidimensional scaling plot based on R_{ST} distances between populations derived from Y-STR data.

Figure 3. BATWING population splitting tree. Numbers on branches show partition posterior probability.

Figure 4. Genome-wide population structure. A) Principal component analysis of ~44,000 SNPs showing the top two components. B) Maximum likelihood tree showing populations relationships.

Figure S1. Y-chromosomal phylogenetic chart. Hierarchical phylogenetic relationships and absolute frequencies of the Y-chromosomal haplogroups observed in Libyan and Moroccan populations. Nomenclature is according to Karafet et al. (2008).

Figure S2. Median joining (MJ) networks. Plotted are MJ networks of Y-STR haplotypes within haplogroups A) E-M78, B) E-M81, C) J-M172, and D) J-M267. The circle sizes are proportional to the haplotype frequencies. The smallest area is equivalent to one individual. Branch lengths are proportional to the number of mutational steps separating two haplotypes.

Figure S3. Inferred population tree with mixture events. A) Tree of population relationships inferred by *TreeMix* allowing six migration events. Horizontal branch lengths are proportional to the amount of genetic drift that has occurred on the branch. B) Residual fit from the maximum likelihood tree. Positive residuals indicate populations where the fit might be improved by adding additional edges.

Figure 1

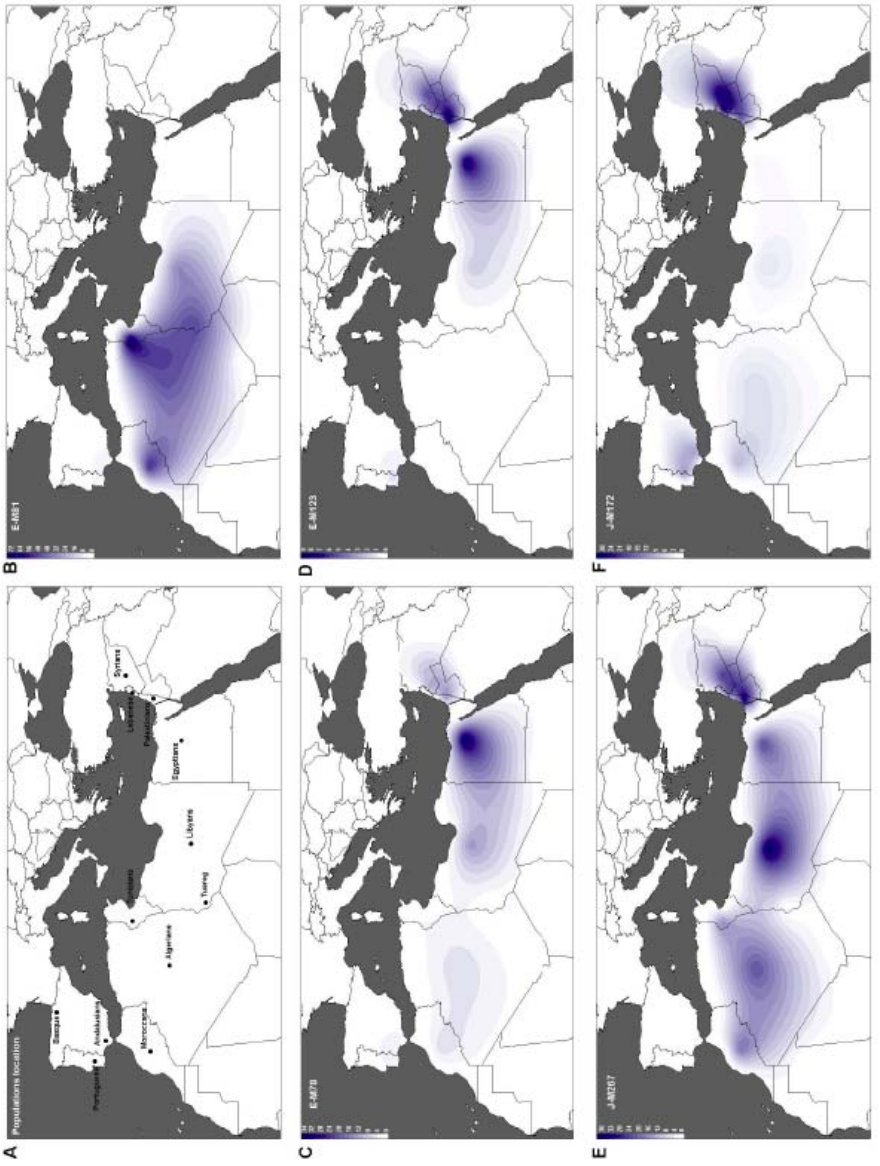
A
Populations location
Basques
Portuguese
Andalusians
Moroccans
Algerians
Tuareg
Libyans
Egyptians
Palestinians
Lebanese
Syrians
B
E-M81
C
E-M78
D
E-M123
E
J-M267
F
J-M172

Figure 2

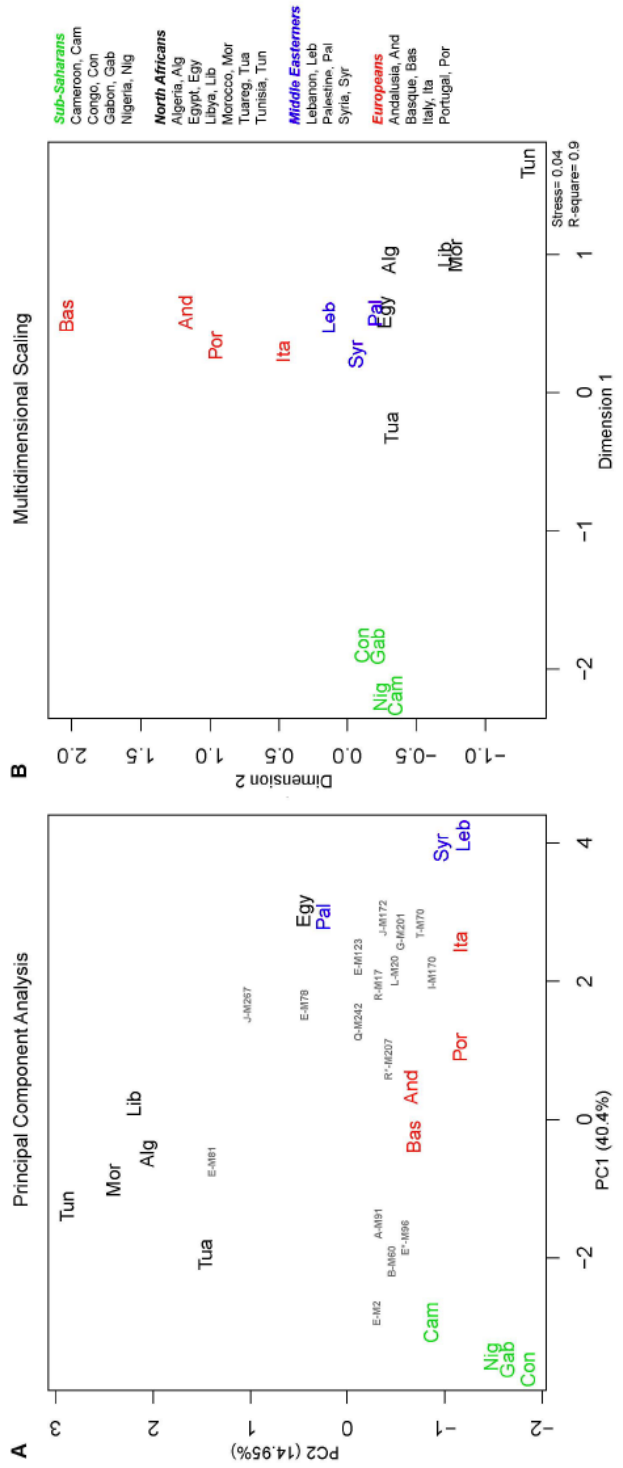
A
Principal Component Analysis
PC2 (14.95%)
PC1 (40.4%)
Tun
Mor
Alg
Lib
Tua
E-M81
E-M2
B-M60
A-M91
E*-M96
Cam
Nig
Gab
Con
Bas
And
Por
Ita
Egy
Pal
Syr
Leb
J-M267
E-M78
Q-M242
R*-M207
R-M17
E-M123
L-M20
J-M172
G-M201
T-M70
I-M170
B
Multidimensional Scaling
Dimension 2
Dimension 1
Stress= 0.04
R-square= 0.9
Sub-Saharans
Cameroon, Cam
Congo, Con
Gabon, Gab
Nigeria, Nig
North Africans
Algeria, Alg
Egypt, Egy
Libya, Lib
Morocco, Mor
Tuareg, Tua
Tunisia, Tun
Middle Easterners
Lebanon, Leb
Palestine, Pal
Syria, Syr
Europeans
Andalusia, And
Basque, Bas
Italy, Ita
Portugal, Por

Figure 3

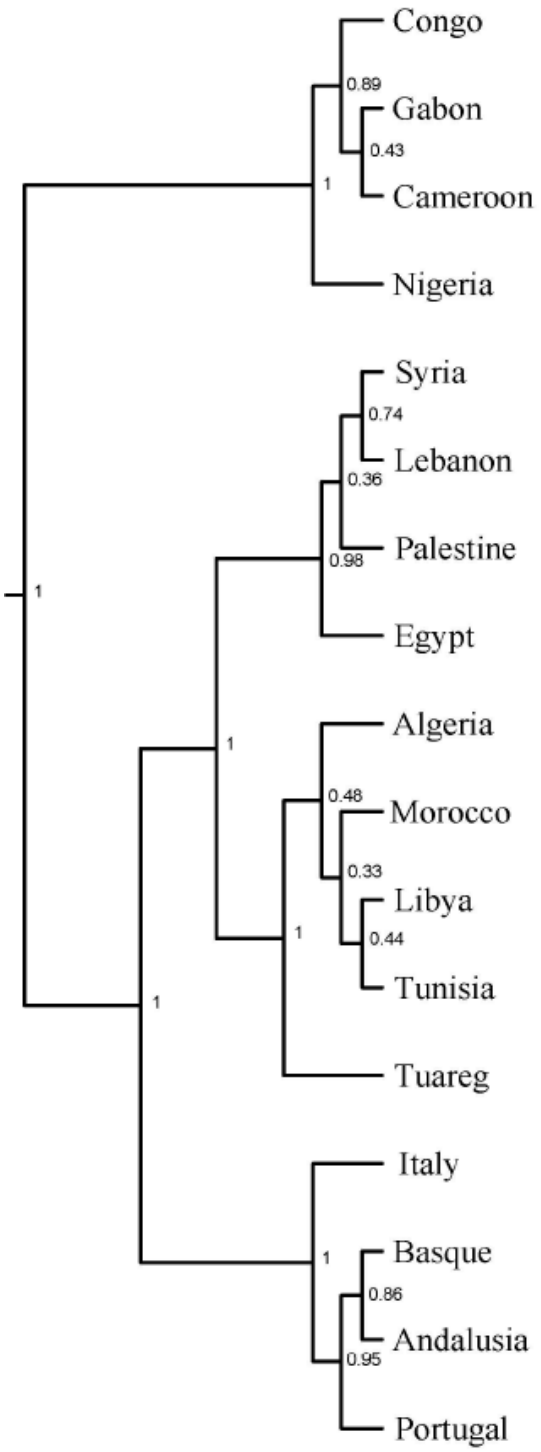
Congo
0.89
Gabon
0.43
Cameroon
1
Nigeria
Syria
0.74
Lebanon
0.36
Palestine
0.98
Egypt
1
Algeria
0.48
Morocco
0.33
Libya
1
0.44
Tunisia
1
Tuareg
Italy
1
Basque
0.86
Andalusia
0.95
Portugal

Figure 4

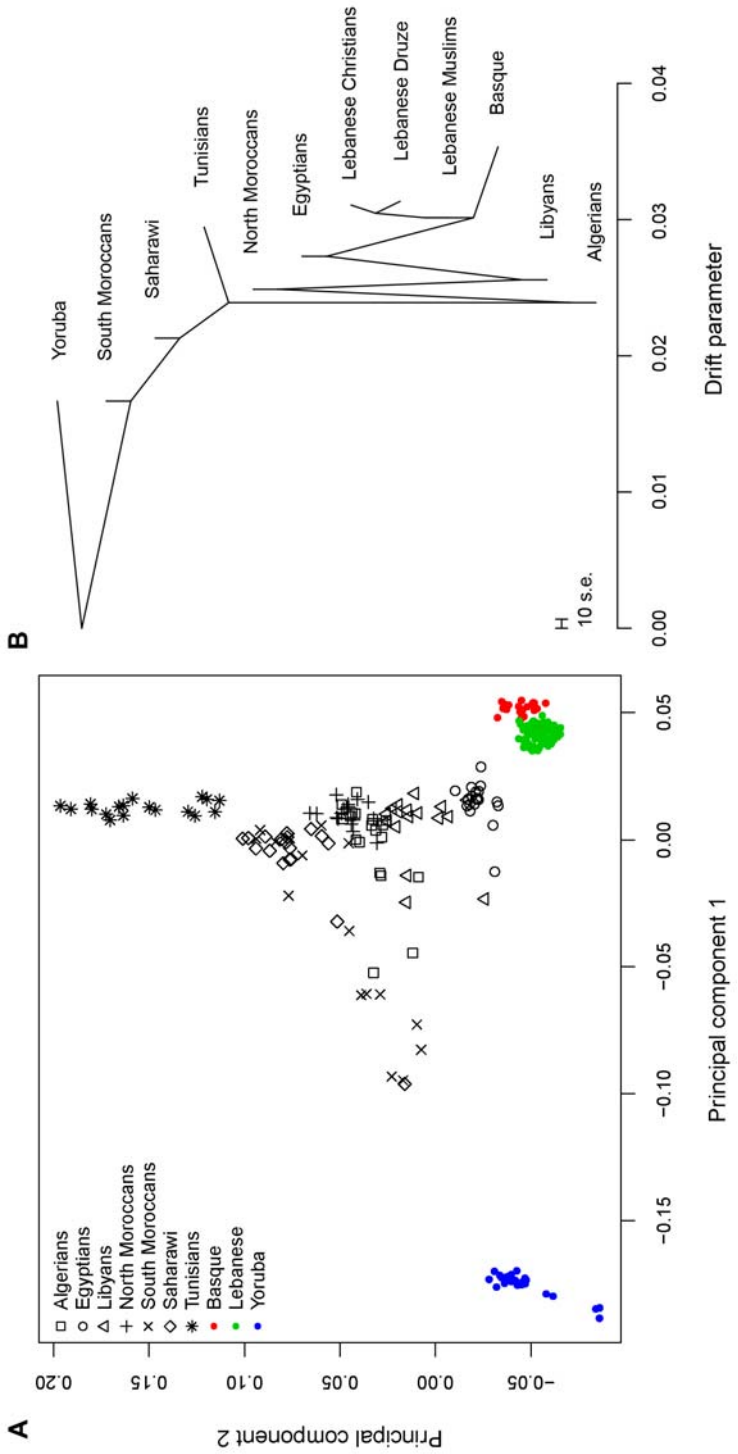
A
B
Principal component 2
Principal component 1
Drift parameter
H
10 s.e.
Algerians
Egyptians
Libyans
North Moroccans
South Moroccans
Saharawi
Tunisians
Basque
Lebanese
Yoruba
Yoruba
South Moroccans
Saharawi
Tunisians
North Moroccans
Egyptians
Lebanese Christians
Lebanese Druze
Lebanese Muslims
Basque
Libyans
Algerians

Figure S1

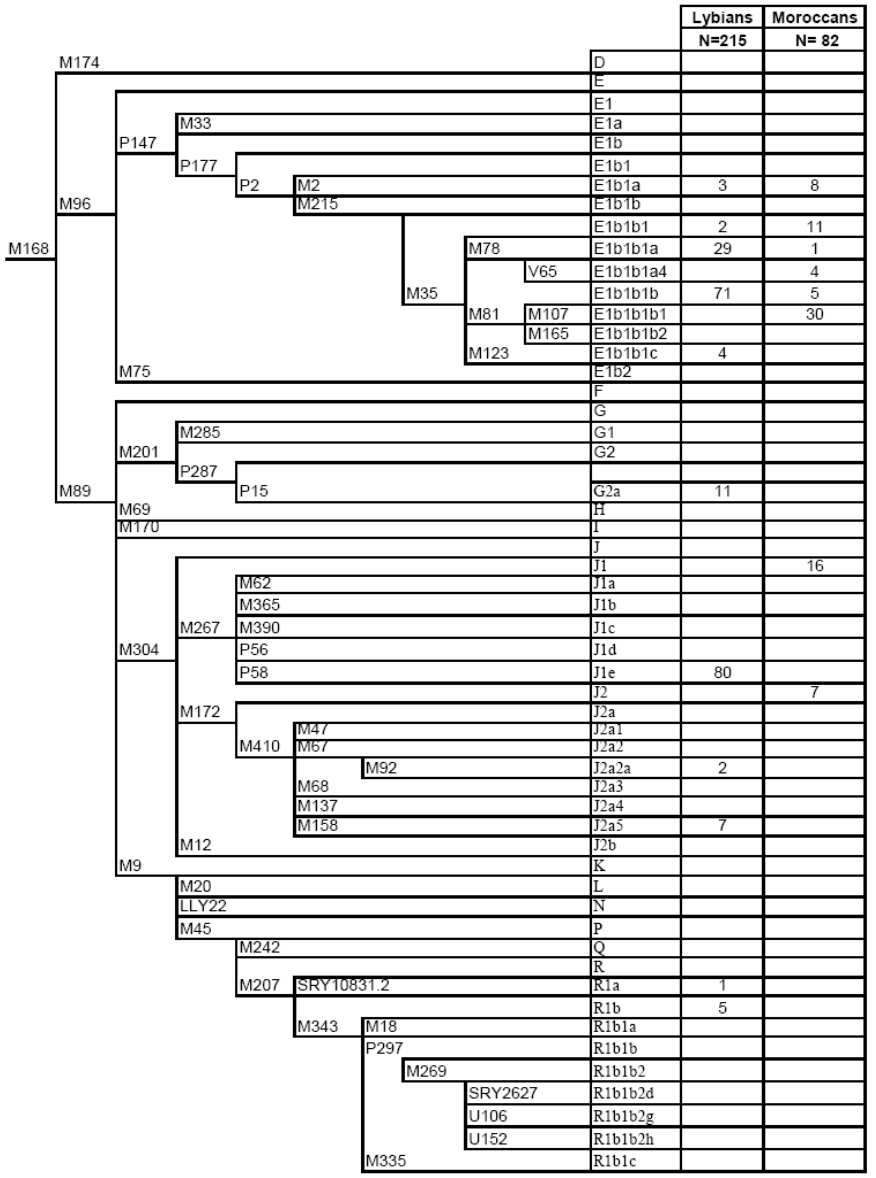
Lybians N=215
Moroccans N= 82
M174
M96
M168
M89
P147
M33
P177
P2
M2
M215
M78
V65
M35
M81
M107
M165
M123
M75
M201
M285
P287
P15
M69
M170
M62
M365
M267
M390
M304
P56
P58
M172
M47
M410
M67
M92
M68
M137
M158
M12
M9
M20
LLY22
M45
M242
M207
SRY10831.2
M343
M18
P297
M269
SRY2627
U106
U152
M335
D
E
E1
E1a
E1b
E1b1
E1b1a 3 8
E1b1b
E1b1b1 2 11
E1b1b1a 29 1
E1b1b1a4 4
E1b1b1b 71 5
E1b1b1b1 30
E1b1b1b2
E1b1b1c 4
E1b2
F
G
G1
G2
G2a 11
H
I
J
J1 16
J1a
J1b
J1c
J1d
J1e 80
J2 7
J2a
J2a1
J2a2
J2a2a 2
J2a3
J2a4
J2a5 7
J2b
K
L
N
P
Q
R
R1a 1
R1b 5
R1b1a
R1b1b
R1b1b2
R1b1b2d
R1b1b2g
R1b1b2h
R1b1c

Figure S2

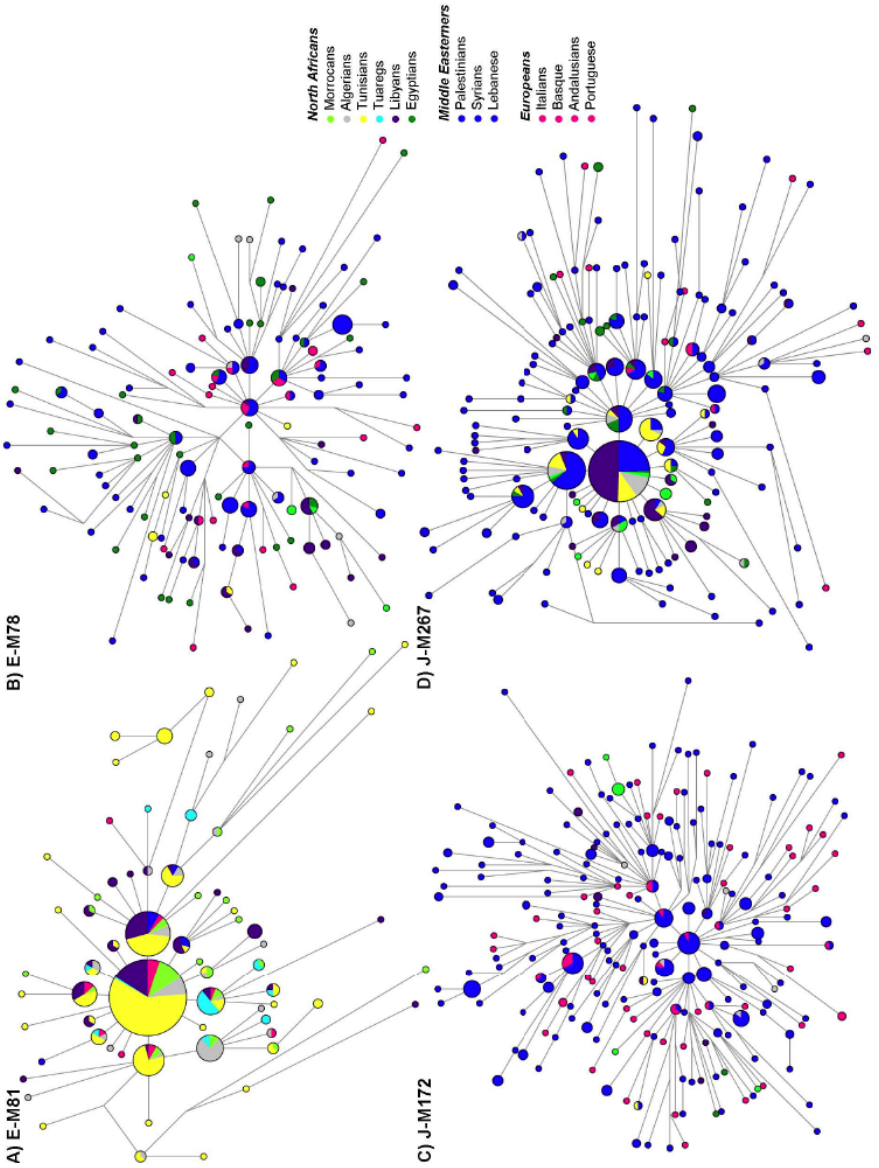
North Africans
Moroccans
Algerians
Tunisians
Tuaregs
Libyans
Egyptians
Middle Easterners
Palestinians
Syrians
Lebanese
Europeans
Italians
Basque
Andalusians
Portuguese
B) E-M78
D) J-M267
A) E-M81
C) J-M172

Figure S3

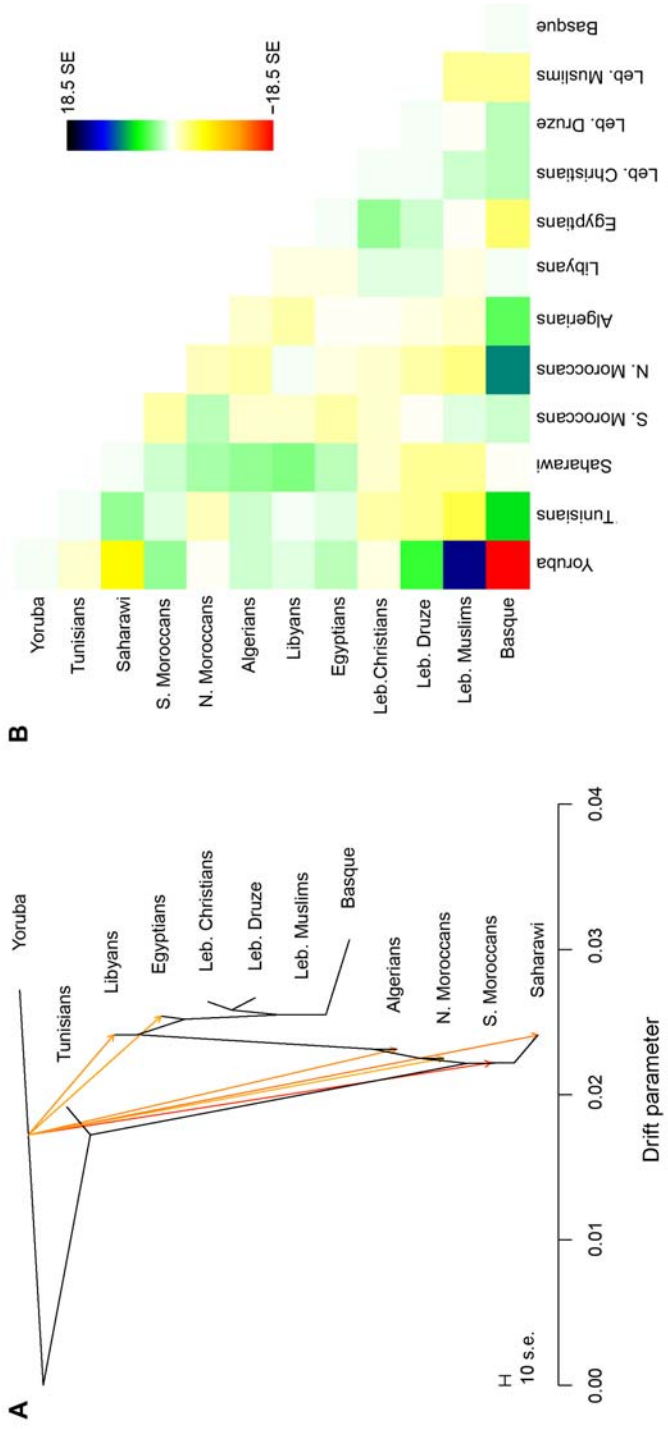
B
18.5 SE
−18.5 SE
Yoruba
Tunisians
Saharawi
S. Moroccans
N. Moroccans
Algerians
Libyans
Egyptians
Leb.Christians
Leb. Druze
Leb. Muslims
Basque
A
H
10 s.e.
0.00
0.01
0.02
0.03
0.04
Drift parameter

Table S1: Populations selected for the Y-chromsome analyzes.

Population	Abbreviation	N	References
North Africa			
Libya	Lib	215	Present study
Morocco	Mor	87	Present study
Tunisia	Tun	239	Fadhlaoui-Zid et al. 2011; Ennafa et al. 2011
Algeria	Alg	102	Robino et al. 2007
Tuareg (Libya)	Tua	47	Ottoni et al. 2011
Egypt	Egy	110	El-Sibai et al. 2009
Europe			
Italy	Ita	162	Onofri et al. 2007
Iberian Peninsula (Andalusians)	And	168	Adams et al. 2008
Portugal	Por	138	Adams et al. 2008
Basque (Spain)	Bas	116	Adams et al. 2008
Middle East			
Lebanon	Leb	577	Zalloua et al. 2008
Syria	Syr	202	Zalloua et al. 2008
Palestine	Pal	367	Zalloua et al. 2008
Sub-Sahara Africa			
Cameroon	Cam	166	Montano et al. 2011
Congo	Con	19	Montano et al. 2011
Gabon	Gab	163	Montano et al. 2011
Nigeria	Nig	132	Montano et al. 2011

Table S2. Y-chromosome haplogroups frequencies in populations selected for this study.

	A-M91	B-M60	E*	E-M123	E-M2	E-M78	E-M81	G-M201	I-M170	J-M172	J-M267	L-M20	Q-M242	R*	R-M17	T-M70
Libya	–	–	0.0076	0.0153	0.0878	0.1107	0.3588	0.042	–	0.0344	0.3053	–	–	0.0343	0.0038	–
Morocco	–	–	–	–	0.0975	0.0610	0.5610	–	–	0.0854	0.1951	–	–	–	–	–
Algeria	–	–	–	–	0.0784	0.0588	0.4510	–	–	0.0490	0.2255	–	0.0098	0.1177	0.0098	–
Tunisia	–	–	0.0043	0.0043	0.0085	0.0213	0.7906	0.0043	–	0.0043	0.1581	–	–	–	0.0043	–
Tuareg	–	–	–	–	0.4255	–	0.4894	–	–	–	–	–	–	0.0213	0.0638	–
Egypt	–	–	–	0.0849	0.0377	0.3585	–	0.0755	–	0.0189	0.2453	0.0189	–	0.0849	0.0283	0.0471
Lebanon	–	–	–	0.0441	0.0068	0.1085	0.0124	0.0678	0.0497	0.2678	0.2079	0.0542	0.0203	0.0859	0.0260	0.0486
Palestine	–	–	–	0.0951	0.0035	0.1549	–	0.0880	0.0458	0.1972	0.3556	0.0071	–	0.0211	0.0106	0.0211
Syria	–	–	–	0.0543	0.0109	0.0815	0.0055	0.0598	0.0489	0.2826	0.2446	0.0543	–	0.0544	0.0815	0.0217
Andalusia	–	–	0.0060	0.0060	–	0.0238	0.0536	0.0357	0.0595	0.1131	0.0238	–	0.0059	0.6488	0.0238	–
Basque	–	–	–	–	–	–	0.0086	–	0.0776	0.0259	0.0086	–	0.0086	0.8707	–	–
Portugal	–	–	0.0362	0.0145	–	–	0.0580	0.1015	0.0290	0.1160	0.0217	–	–	0.5145	0.0217	0.0362
Italy	–	–	–	0.0875	–	0.1000	–	0.1250	0.0812	0.1563	0.0563	–	–	0.3625	0.0062	0.0250
Cameroun	–	0.0546	0.0121	–	0.9212	–	–	–	–	–	–	–	–	0.0121	–	–
Congo	–	0.0526	0.1053	–	0.8421	–	–	–	–	–	–	–	–	–	–	–
Gabon	–	0.0797	0.0552	–	0.8221	–	–	–	–	–	–	–	–	0.0430	–	–
Nigeria	0.0303	0.0379	0.0303	–	0.8864	–	–	–	–	–	–	–	–	0.0151	–	–

Table S3. Analyses of Molecular Variance (AMOVA) in North African and Middle Eastern samples based on Y-STR haplotypes and Y-SNP haplogroups. Acronyms are listed in Table S1.

Groups	Among groups		Among populations within groups		Within populations	
	Y-STR	Y-SNP	Y-STR	Y-SNP	Y-STR	Y-SNP
All populations			11.15***	15.17***	88.85***	84.83***
(Mor) vs (Alg, Tun, Tua, Lib , Egy, Leb, Syr, Pal)	-1.35ns	0.70ns	11.41***	15.01***	89.94***	84.30***
(Mor, Alg) vs (Tun, Tua, Lib, Egy, Leb, Syr, Pal)	-0.49ns	2.78ns	11.29***	14.27***	89.20***	82.95***
(Mor, Alg, Tun) vs (Tua, Lib, Egy, Leb, Syr, Pal)	12.67*	18.67*	5.05***	5.91***	82.29***	75.42***
(Mor, Alg, Tun, Tua) vs (Lib, Egy, Leb, Syr, Pal)	10.72*	19.57**	5.68***	4.91***	83.61***	75.52***
(Mor, Alg, Tun, Tua, Lib) vs (Egy, Leb, Syr, Pal)	10.78*	16.63*	4.60***	4.75***	84.62***	78.62***
(Mor, Alg, Tun, Tua, Lib, Egy) vs (Leb, Syr, Pal)	8.86ns	12.34ns	5.52***	7.02***	85.63***	80.64***
(Mor, Alg, Tun,) vs (Tua, Lib, Egy)	0.78ns	8.33 ns	3.92***	11.43***	95.30***	80.23***
(Mor, Alg, Tun,) vs (Tua, Lib, Egy) vs (Leb, Syr, Pal)	10.13**	14.66**	3.77***	4.40***	86.10***	80.93***

*** P<0.0001; ** P<0.01; * P<0.05; ns: not significant

Table S4. BATWING results showing times of demographic factors for Y-chromosomes from North Africans

Population 1	Population 2	TMRCA	Growth start	Split
Algerians	Egyptians	15 (11-25)†	8.9 (6-12.5)	2.8 (2.1-3.6)
Algerians	Libyans	15 (10.5-26)	11 (6.5-18)	1.3 (0.9-1.9)
Algerians	Tuareg	15.5 (10.5-27)	7.5 (2.5-14.5)	1.1 (0.5-2.3)
Algerians	Moroccans	22 (13.5-46)	1.2 (0.8-2.1)	1.2 (0.7-1.6)
Algerians	Tunisians	32 (17-74.5)	1.8 (1.1-2.7)	1.3 (0.9-1.7)
Egyptians	Libyans	14.5 (10.5-22.5)	10 (7-14)	1.9 (1.4-2.5)
Egyptians	Tuareg	16 (11.5-25)	11 (8-15.5)	2.8 (1.3-4.6)
Egyptians	Moroccans	15 (10.5-24)	9.5 (6.7-13.2)	2.7 (2-3.6)
Egyptians	Tunisians	15 (11-32)	10 (7.3-13.5)	2.6 (2-3.4)
Libyans	Tuareg	16.5 (11-27)	13 (7.5-22.5)	1.9 (1.1-3.2)
Libyans	Moroccans	14.5 (10-25.5)	12 (3.7-19.5)	1.3 (0.8-1.9)
Libyans	Tunisians	30.5 (14-79)	10.7 (0.6-24)	1 (0.7-1.4)
Moroccans	Tunisians	42 (20.5-96)	1.4 (0.6-7.5)	1.2 (0.8-1.8)
Tuareg	Moroccans	15 (10-26)	8.5 (2.3-18)	1.6 (0.7-3.2)
Tuareg	Tunisians	44.5 (21-98.5)	1.2 (0.2-25.5)	1.7 (1-2.9)

† Median value with 95%CI in thousand of years ago

Table S5. 3-population test showing gene flow to North Africans

Target	Source1	Source2	Minimum[a] *f3*	S.E.	Z-score
Algerians	Yoruba	Basque	-0.00856386[b]	0.000216115	-39.6263
Algerians	Yoruba	Lebanese Christians	-0.00712749	0.000181691	-39.2285
Egyptians	Yoruba	Basque	-0.0064804	0.00020486	-31.6333
Egyptians	Yoruba	Lebanese Christians	-0.00573643	0.000178883	-32.0681
Libyans	Yoruba	Basque	-0.00777087	0.00019491	-39.8689
Libyans	Yoruba	Lebanese Christians	-0.00664719	0.000161584	-41.1377
North Moroccans	Yoruba	Basque	-0.00657986	0.00019616	-33.5433
North Moroccans	Yoruba	Lebanese Christians	-0.00475158	0.000160078	-29.683
South Moroccans	Yoruba	Basque	-0.0111038	0.000202255	-54.9001
South Moroccans	Yoruba	Lebanese Christians	-0.00977224	0.000175436	-55.7025
Saharawi	Yoruba	Basque	-0.00677832	0.000228239	-29.6983
Saharawi	Yoruba	Lebanese Christians	-0.00557365	0.000200604	-27.7843
Tunisian Berbers	Yoruba	Basque	0.000352494[c]	0.000275206	1.28084
Tunisian Berbers	Yoruba	Lebanese Christians	0.00209835	0.000252313	8.31648

[a] The table only lists the two lowest f3 statistics observed in North Africans.
[b] A significantly negative value of the f3 statistic implies that target population is admixed.
[c] A positive f3 suggests that target population is unadmixed.

4. Discussion

The current work investigates the genetic diversity of human populations in a vast geographic area extending from Central Asia to North Africa. We use variations on the human genome (uniparental and genome-wide) and employ various analysis methodologies to infer on the origin and history of these populations.

In addition to the discussions presented with each study in the previous chapter, the findings of this work will be discussed in this chapter further with interpretations significant to the general understanding of modern humans' genetic diversity.

4.1 Inferences on past demographic processes

4.1.1 Emergence of modern human populations

The three major populations studied in this thesis; Afghans, Levantines, and North Africans, show that major population structures emerged at the end of the last ice age. During this time, previously isolated human populations in glacial refugia started expanding to new territories and admixing with one another. This arising new genetic diversity has probably been limited by geographical distances forming the major regional population stratifications observed in humans today (Figure 11).

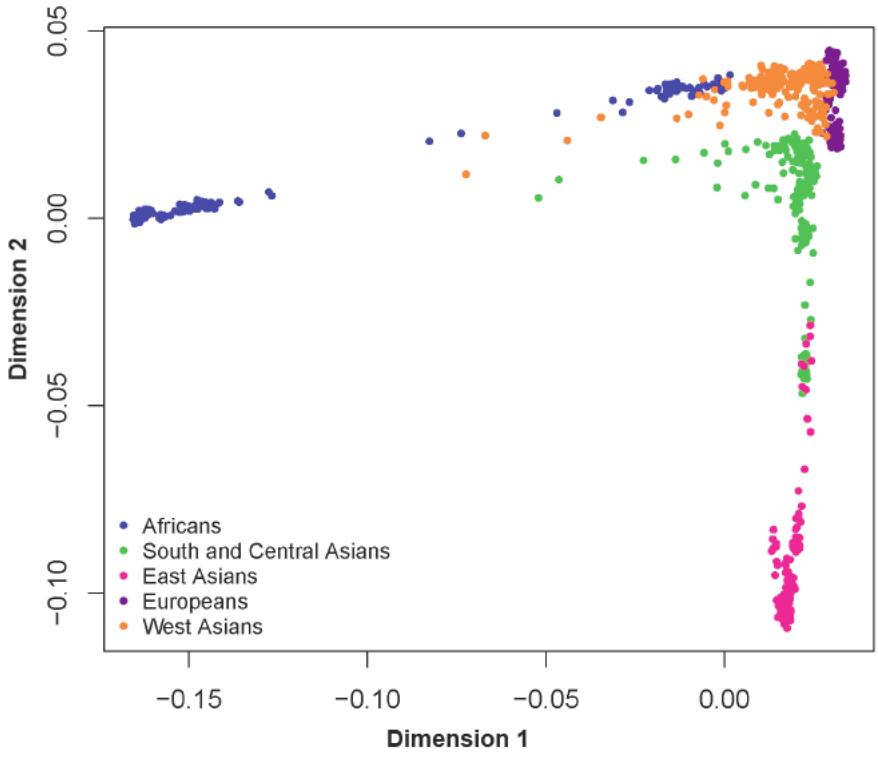

Figure 11. Multidimensional scaling of >240K SNPs showing the top two dimensions. Plot shows global diversity using 50 populations classified in 5 regional groups. Produced with data from (Li *et al.*, 2008, Behar *et al.*, 2010, Haber *et al.*, 2013).

Our coalescence time estimate for the paternal lineages in North Africa is ~15,000 ya for most populations. These dates coincide with major environmental changes in North Africa following the full glacial hyperarid conditions during the Last Glacial Maxima (LGM). Humid conditions started in North Africa ~14,500 ya transforming the area into a verdant landscape vegetated with annual grasses and shrubs which attracted hunter-gatherers who spread into the region (Bar-Yosef, 1987, Brovkin & Claussen, 2008, Kropelin *et al.*, 2008). Similarly, our time estimate of divergence between the Levantine and European ancestral populations is

~15,900-9,100 ya suggesting that population migration to Europe from the Near East could have started after the LGM warming and continued until the Neolithic. Also, our results indicate that the Afghan population split from Iranians, Indians and East Europeans at about 10,600 ya (95% CI 7,100–15,825), which also coincides with the warming period following the LGM.

4.1.2 Development of fine population structures

Climate change and geography were major factors shaping humans' genetic diversity and forming major regional population structures. However, these regional populations were also subjected to other diverse processes leading to further population stratifications. This section will summarize these processes and how they have influenced the diversity of the populations studied in this work.

Role of the cultural systems

Recent cultural developments, such as the founding and expansion of religions and ethnicities, have had a strong impact on the genome of modern populations. When cultural affiliation dictates the selection of mates in a population, different groups become genetically separated and drift fixes different alleles in each group creating population structures.

Our data shows that population structures in Afghanistan today are highly correlated with ethnicity. We found substantial differences among the various groups of Afghanistan. In addition, we found that the flow of paternal lineages among the various ethnic groups is

very limited, and it is consistent with high level of endogamy practiced by these groups. Afghanistan's harsh geography of mountains, deserts and steppes, could have facilitated the establishment of social organizations within expanding populations, and helped maintaining genetic boundaries among groups that have developed over time into distinct ethnicities.

Similarly, our results show recent genetic stratifications in the Levant are driven by the religious affiliations of the populations within the region. For example, we show that cultural transition in the region to Islam appears to have introduced major rearrangements in population relations through interaction and admixture with culturally similar but geographically remote populations. This is revealed through the co-ancestry matrix constructed from genome-wide haplotypes that show Muslim populations of Syrians, Palestinians and Jordanians cluster with other Muslim populations as distant as Morocco (~4,000 Km away) but not with non-Muslim populations, like the Druze, Jews and Christians, that probably live in the same or neighboring village/city.

Migration and gene flow

Migration to inhabited regions leads to admixture between populations and is consequently a major force shaping genetic diversity. For example, we have found that expanding populations into Afghanistan have been admixing with the indigenous populations, giving the Afghans distinctive genetics from the expanding source and from their surroundings. This is evident in the

Afghan Hazara and Afghan Uzbek who have assimilated expanding Mongols and Turco-Mongols and now have at least third to half of their paternal lineages of East Asian origin. On the other hand, gene flow to Afghanistan from India marked by Indian lineages, L-M20, H-M69, and R2a-M124, seems to mostly involve Pashtuns and Tajiks, creating an Afghan-Indian population structure that excludes the Hazaras and Uzbeks.

Admixture has also profoundly shaped the genomes of North Africans. Our formal admixture tests, using genome-wide SNPs, show that most North Africans today are a mixture of populations related to current Africans and Eurasians which concur with previous findings reporting a ‘‘back-to-Africa’’ gene flow more than 12,000 ya (Henn *et al.*, 2012a).

In the Levant region we have used admixture patterns to infer on historical events. We have noticed from the coancestry matrix (constructed from genome-wide haplotypes) that the gene flow from the Middle East to the Levant region was accompanied by an introgression of sub-Saharan haplotype chunks in the Levantines.

This provided an opportunity to use the rate of exponential decline of LD on the African haplotypes to date the admixture events in the Levant. The African haplotypes are useful in our case since African populations provide an unadmixed (relative to the Middle Easterners for example) reference to test LD decay. Interestingly, our estimation of admixture dates correlated well with documented historical events. For example, we have estimated that the Druze have stopped admixing with surrounding populations around 1,275-

1,025 ya which correlates with the emergence of the Druze faith in 986 CE.

Rise and fall of civilizations

The formation of the first urban civilizations accompanied by the development of trade and commerce has probably brought diverse populations from different regions into contact, shaping the genetic diversity of these populations. For example, we have found that the first genetic structure between the different Afghan groups started 4,700 ya during the Bronze Age and the formation of the first civilizations in the region. It is interesting that the economy of these civilizations was most likely supported by the agricultural ancestral populations which we found remained unstructured for about 6,000 years. When urbanization developed, populations splits accelerated. Civilization decline can also influence genetic diversity. We show that Egypt splits from other North Africans ~2,800 ya when the kingdom declined in power and was conquered by Assyrians and Persians. The conquest of Egypt has probably limited gene flow from other North Africans and increased gene flow from the Near East, creating the split with North Africa.

4.2 Overview of used methodologies

We have used various markers and methods to investigate human genetic diversity in this work. This section summarizes the usefulness of these methodologies by assessing their strength and weaknesses in inferring on demographic processes.

4.2.1 DNA markers

Markers used in this work can be classified into two categories: variations on the Y-chromosome and variations on the whole genome. The inheritance pattern of these two types of markers is different and consequently the methodologies and inferences are also different.

Genetic relationships between human groups have been heavily studied using uniparental (Y-chromosome and mtDNA) markers. Using these markers involves phylogeographic analyses based on highly detailed and regional specific haplogroup trees (Figure 12), consequently allowing very fast inferences on the ancestry of a sample. For example, we can trace back haplogroup C3 found in Afghan Hazara to a Mongol origin and trace back haplogroup R2 found in Pashtun to an Indian origin.

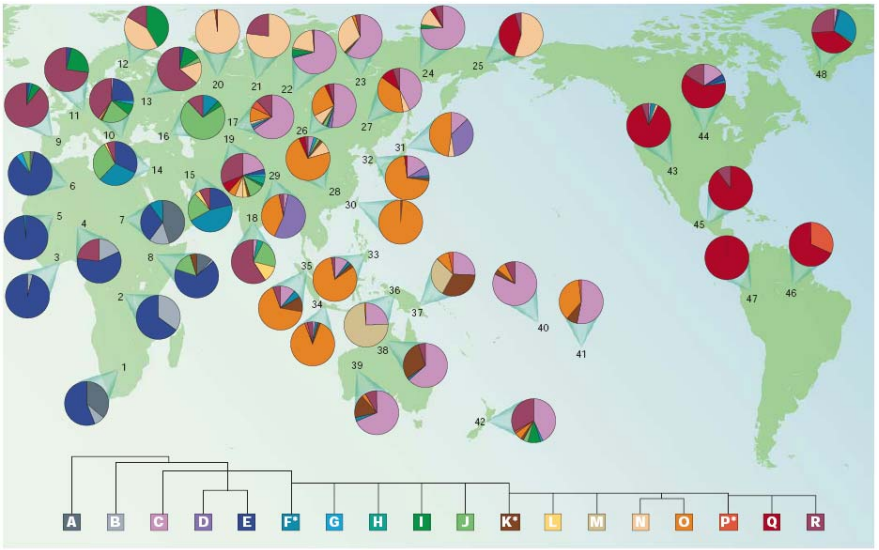

Figure 12. Global distribution of Y haplogroups. Phylogeographic analyses based on these haplogroups allow fast inference of ancestry. From (Jobling & Tyler-Smith, 2003).

The Y-chromosome provide a powerful tool to study populations, however it is highly susceptible to selection and drift. In addition, the ancestry details we get from a single locus are limited. For example, we can trace haplogroup R1b to Europe but additional details become harder to extract even with heavy sub-genotyping. For instance, haplogroup R1b1a2a1a2b1a1 is found in Spain, Italy, Belgium, UK, Denmark, Germany, and Hungary.

Evidently, the problem of resolution is solved when using thousands of markers spread on the whole genome. The early studies that investigated patterns of human genetic variation using thousands of genome-wide markers were fascinatingly surprising. It showed that an individual's geographic origin can be traced back accurately to within a few hundred kilometers (Figure 13).

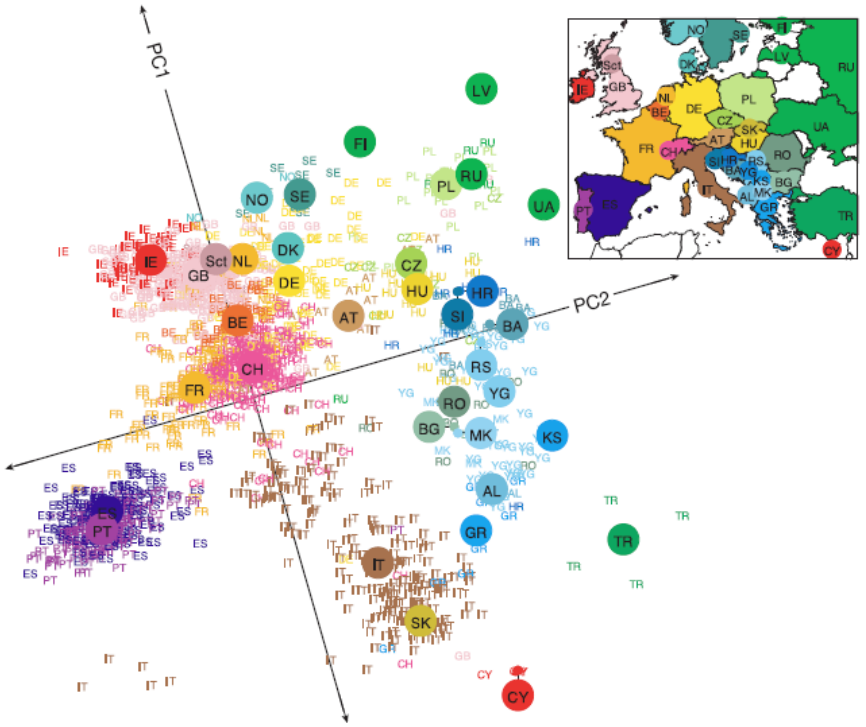

Figure 13. Population structure within Europe. Using thousands of genome-wide SNPs can potentially pinpoint an individual's geographical origin. From (Novembre *et al.*, 2008).

The genome-wide SNPs used in this thesis were derived from a genotyping array. One potential problem with using genotyping arrays is ascertainment bias which is caused by the way markers are chosen. A high proportion of SNPs on these arrays were discovered in European populations. In addition, the arrays were designed with GWAS in mind and therefore have regions previously implicated with diseases enriched with SNPs. This type of bias has the potential of distorting population genetic inferences and can only be completely solved by using whole-genome sequencing. However,

the effect of ascertainment bias on the results of this work have probably been reduced for two reasons: 1- The Levantine population is very close genetically to Europeans, having the smallest F_{st} (0.035) compared to other worldwide populations. Therefore, the Levantines potentially catch the genetic diversity as in the populations in which most of the array SNPs were discovered.

2- We made genetic inferences from haplotype data which reduces ascertainment bias because haplotypes are more likely to be polymorphic in multiple populations. Haplotype statistics are therefore less affected by ascertainment bias than single locus statistics (Lachance & Tishkoff, 2013, Novembre & Ramachandran, 2011).

Although the emergence of SNP array technology five years ago appeared to replace the use of uniparental markers for inferences of population structure, recent studies show that these markers can still hold very interesting information on our past. For example, a recent study discovered an African American Y-chromosome with an estimated TMRCA of 338,000 ya (95% confidence interval = 237,000-581,000 ya), exceeding the age of the oldest anatomically modern human fossils and suggesting ancient population structure and the possibility of archaic introgression of Y-chromosomes into anatomically modern humans (Mendez *et al.*, 2013). Therefore, DNA markers of future studies on human populations will certainly be variations derived from whole-genome sequencing, including whole Y-chromosome and mtDNA genome sequencing.

4.2.2 Clustering methods

The first analysis to explore our populations, using any type of DNA markers, was plotting allele frequencies (PCA) or genetic distances (MDS) in two dimensions. This provided very fast assessment of the data and the populations and allowed rapid inferences on the relation of the studied populations. However, the information we get from these plots is limited and inferences are usually from visual inspection of the distribution of individuals or populations on a plot that catches only part of the whole diversity. This type of inferences could be problematic in populations with fine genetic structure as is the case in the Levant region. We found that using an approach such as constructing a coancestry matrix using ChromoPainter (Lawson *et al.*, 2012) and visualizing genetic relationships using a Maximum likelihood tree (in our case constructed from shared haplotype chunks), efficiently capture information on fine population structure. However, this approach does not allow for the possibility of migrations between groups, and therefore inferences on the history of highly admixed populations, such as the North Africans studied here, can be limited. Therefore, we used a model that improve simple bifurcating trees by allowing for both population splits and gene flow (*TreeMix*) (Pickrell & Pritchard, 2012) showing that North Africans diversity can be explained by migrations from Sub-Saharan Africans and Near Easterners.

Another clustering method used in this work is ADMIXTURE (Alexander *et al.*, 2009), a model based approach to classify individuals to hypothetical ancestral populations. One problem of

ADMIXTURE is that it requires a prior specification of the number of ancestral populations *K* and there is no effective method to predict the "truth" *K*. However, an advantage of ADMIXTURE is that it can identify ancestral components and distances between these components, independent of subsequent admixture events. Therefore, in this work we use the ChromoPainter approach discussed above to identify fine populations subdivisions without the drawback of specifying a *K* value, and use ADMIXTURE to estimate the genetic distances between the ancestral components independent of subsequent admixture events. For inference of admixture we find that formal tests of admixture, as in the 3-population test (Patterson *et al.*, 2012), provide very useful and unambiguous evidence of admixture even if the gene flow events occurred hundreds of generations ago, such as in the case of gene flow from the Near East to North Africa.

4.3 Significance to medical studies

The aim of this work was to study the genetic diversity of populations to learn about their origin and history. However, studies like the one presented in this thesis, can also be very important to medical genetics since genome-wide association studies (GWAS) can be confounded by population stratification. The most common approach to correct for population stratification is to include the principal components from the PCA as covariates. However, there is no solid method to determine how many components to include and selection is usually subjective with inconclusive results. Correction for stratification becomes even more complicated in populations like the Levantines where genetic structure is recent and not driven by the usual factors of distance and geography.

The current study can be useful to GWAS at two stages:

1- *At the design stage*. Having an insight on the factors driving population structure is crucial for GWAS. For example, the demographic information usually collected for GWAS from patients includes the place of origin such as country/town/city/village. Since we have identified that religion rather than geography is correlated with genetic structure in the Levant, collecting religion information from patients (which is unusual) becomes more important from knowing their place of origin.

2- *At the analysis stage*. Our study has identified ancestry related markers among the different Levantine groups, removing these SNPs will probably remove most stratification. However, there are many SNP arrays available with different coverage and our previously identified SNPs will not be able to account for all

ancestry differences. Therefore, we have developed a simple pipeline (Figure 14) that makes use of previously identified ancestry informative SNPs and well classified subjects and can be applied to any GWAS dataset irrelevant of the genotyping array. The method makes use of available 1000 Genomes Project reference panels to impute the resolved ancestry dataset as well as the GWAS dataset to the same SNP coverage resolution and then merge the two datasets. We next use a model-based clustering method to classify subjects into groups and label the groups based on the membership of the resolved ancestry samples. We then extract the ancestry related SNPs between the groups and exclude it from the GWAS analysis.

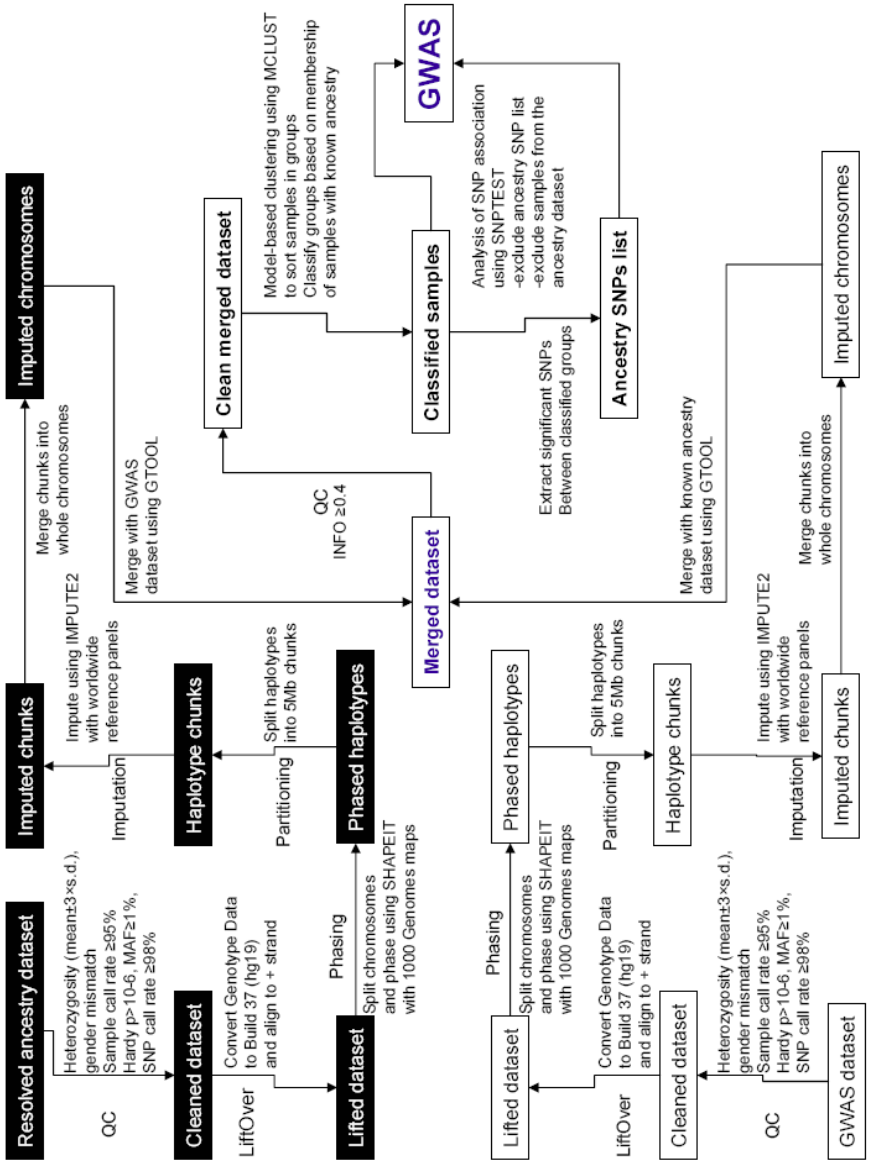

Figure 14. Pipeline for correcting population structure in GWAS. Pipeline starts simultaneously from bottom left and bottom right. Method uses datasets from ancestry studies to avoid type 1 errors from population stratification in GWAS.

We asses the efficiency of this method by applying it to an unpublished dataset of 3290 Lebanese subjects with phenotype information on Type-2 diabetes mellitus (T2D). The samples were genotyped on a HumanOmniExpress BeadChip which provides an opportunity to test how well the pipeline performs when the GWAS array is different from our ancestry study array (Human 660 and 610 Quad BeadChips). Imputation results show that the HumanOmniExpress (~700,000 SNPs) has slightly better imputation accuracy compared to the Quad BeadChips (~550,000 SNPs) (Figure 15).

We conducted the association study for T2D in the Lebanese population after excluding the list of ancestry correlated SNPs identified in the pipeline. We found that most significant SNPs were in regions of the *TCF7L2* and *CDKAL1* genes (Figure 15). These two genes have been previously implicated in T2D in other populations (Steinthorsdottir *et al.*, 2007, Zeggini *et al.*, 2007, Sladek *et al.*, 2007).

This analysis shows that studies on the demographic history of populations can be very valuable to medical genetics especially by giving insights on population stratifications; a bugbear for genome-wide association studies.

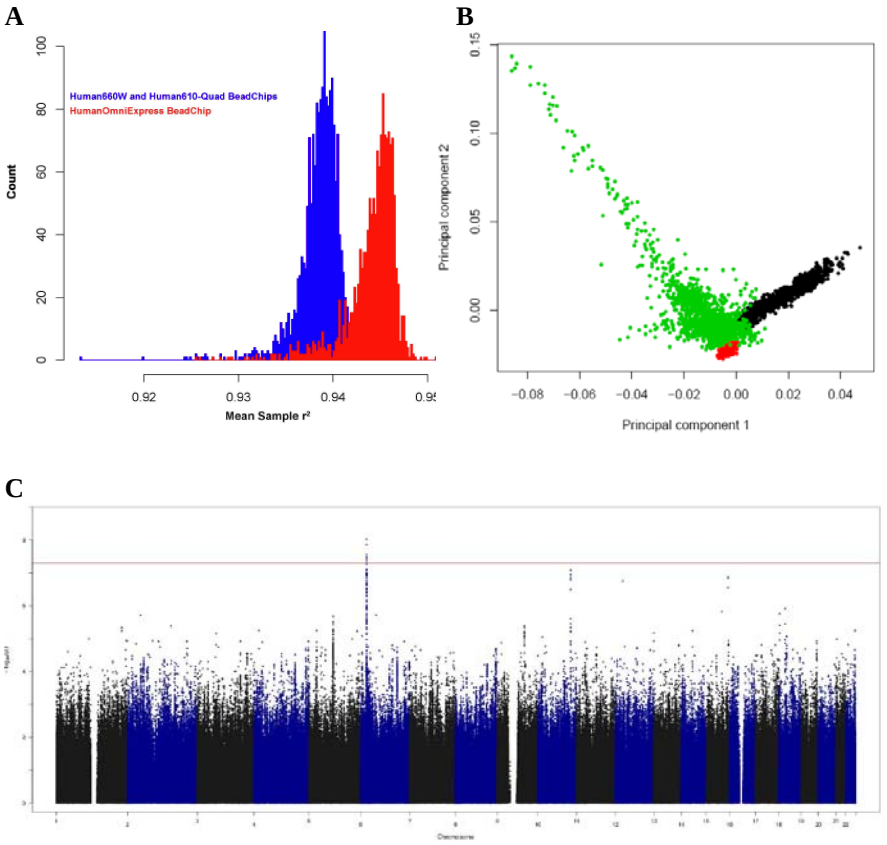

Figure 15. Results of the GWAS on Lebanese with Type 2 diabetes. A) Per-sample imputation accuracy measured by r^2 between true genotypes and genotypes predicted by imputation, averaged over imputation chunks. Accuracy increases with increased array SNP coverage. B) PCA with predicted classification of Lebanese into three groups. Green: Muslims; black: Christians; red: Druze. C) Manhattan plot showing results of a GWA analysis in 3,290 Lebanese patients for >5,000,000 genotyped and imputed SNPs. The Y-axis corresponds to the significance of the association (-log10 p-values). The X-axis represents the physical location of the variant colored by chromosome.

4.4 Concluding remarks

Rapid technological advancements have made genetic data available to scientists to apply evolutionary concepts to a wide range of problems that we could not have imagined possible just 15 years ago. In light of recent advances in many areas of biology and the continuous flow of genetic data, Theodosius Dobzhansky's essay "Nothing in biology makes sense except in the light of evolution" is being cited more than ever. Although the essay itself is probably more cited than read (Dobzhansky argued that God used evolution to produce the diversity of life), the enthusiasm to Dobzhansky's dictum signals an emerging scientific and social embracement of evolution as the main interpreter of the biological world around us.

Evolutionary approaches have helped reconstruct the origin and history of humans, as shown in this thesis through studying the genomes of Central Asians, Near Easterners, and North Africans. We show that many aspects of modern human diversity can be explained by developments during the warming period after the last ice age. The favorable climate conditions and the spread of agriculture resulted in a human population explosion from few millions to seven billions in less than 10,000 years, triggering an extraordinary evolutionary trial and error on the genome of modern humans. This gives us a unique opportunity to understand how evolution shapes the diversity of populations and enables our understanding of the kinds of mechanisms that can affect the future evolutionary trajectory of humans.

5. Bibliography

Alexander, D.H., Novembre, J. & Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res,* 19**,** 1655-64.

Armitage, S.J., Jasim, S.A., Marks, A.E., Parker, A.G., Usik, V.I. & Uerpmann, H.P. (2011) The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia. *Science,* 331**,** 453-6.

Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P.F., Morrow, B., Friedman, E., Oddoux, C., Burns, E. & Ostrer, H. (2010) Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. *Am J Hum Genet,* 86**,** 850-9.

Balter, M. (2011) Was North Africa the launch pad for modern human migrations? *Science,* 331**,** 20-3.

Bar-Yosef, O. (1987) Pleistocene Connexions between Africa and Southwest Asia: An Archaeological Perspective. *The African Archaeological Review,* 5**,** 29-38.

Bar-Yosef, O. (1992) The role of western Asia in modern human origins. *Philos Trans R Soc Lond B Biol Sci,* 337**,** 193-200.

Behar, D.M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., Khusnutdinova, E.K., Balanovsky, O., Semino, O., Pereira, L., Comas, D., Gurwitz, D., Bonne-Tamir, B., Parfitt, T., Hammer, M.F., Skorecki, K. & Villems, R. (2010) The genome-wide structure of the Jewish people. *Nature,* 466**,** 238-42.

Berti, A., Barni, F., Virgili, A., Iacovacci, G., Franchi, C., Rapone, C., Di Carlo, A., Oddo, C.M. & Lago, G. (2005) Autosomal STR frequencies in Afghanistan population. *J Forensic Sci,* 50**,** 1494-6.

Brovkin, V. & Claussen, M. (2008) Comment on "Climate-driven ecosystem succession in the Sahara: the past 6000 years". *Science,* 322**,** 1326; author reply 1326.

Bustamante, C.D. & Henn, B.M. (2010) Human origins: Shadows of early migrations. *Nature,* 468**,** 1044-5.

Cavalli-Sforza, L.L. & Edwards, A.W. (1967) Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet,* 19**,** 233-57.

Degiorgio, M., Jakobsson, M. & Rosenberg, N.A. (2009) Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc Natl Acad Sci U S A,* 106**,** 16057-62.

Di Cristofaro, J., Buhler, S., Temori, S.A. & Chiaroni, J. (2011) Genetic data of 15 STR loci in five populations from Afghanistan. *Forensic Sci Int Genet*.

Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol,* 29**,** 1969-73.

Dupree, L. (1964) Prehistoric Archeological Surveys and Excavations in Afghanistan: 1959-1960 and 1961-1963. *Science,* 146**,** 638-40.

Dupree, L. (1980) *Afghanistan.* Princeton: Princeton University Press.

Excoffier, L., Smouse, P.E. & Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics,* 131**,** 479-91.

Felsenstein, J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics,* 5**,** 164–166.

Fernandes, V., Alshamali, F., Alves, M., Costa, M.D., Pereira, J.B., Silva, N.M., Cherni, L., Harich, N., Cerny, V., Soares, P., Richards, M.B. & Pereira, L. (2012) The Arabian cradle: mitochondrial relicts of the first steps along the southern route out of Africa. *Am J Hum Genet,* 90**,** 347-55.

Fisher, R.A. (1930) *The genetical theory of natural selection.* Oxford,: The Clarendon press.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., De La Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. & Paabo, S. (2010) A draft sequence of the Neandertal genome. *Science,* 328**,** 710-22.

Haber, M., Gauguier, D., Youhanna, S., Patterson, N., Moorjani, P., Botigue, L.R., Platt, D.E., Matisoo-Smith, E., Soria-Hernanz, D.F., Wells, R.S., Bertranpetit, J., Tyler-Smith, C., Comas, D. & Zalloua, P.A. (2013) Genome-wide diversity in the levant reveals recent structuring by culture. *PLoS Genet,* 9**,** e1003316.

Henn, B.M., Botigue, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., Fadhlaoui-Zid, K., Zalloua, P.A., Moreno-Estrada, A., Bertranpetit, J., Bustamante, C.D. & Comas, D. (2012a) Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet,* 8**,** e1002397.

Henn, B.M., Cavalli-Sforza, L.L. & Feldman, M.W. (2012b) The great human expansion. *Proc Natl Acad Sci U S A,* 109**,** 17758-64.

International Hapmap Consortium (2005) A haplotype map of the human genome. *Nature,* 437**,** 1299-320.

Jobling, M., Hollox, E., Hurles, M., Kivisild, T. & Tyler-Smith, C. (2013) *Human evolutionary genetics.* New York: Garland Science.

Jobling, M.A. & Tyler-Smith, C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet,* 4**,** 598-612.

Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature,* 217**,** 624-6.

Kimura, M. & Crow, J.F. (1964) The Number of Alleles That Can Be Maintained in a Finite Population. *Genetics,* 49**,** 725-38.

Kimura, M. & Ohta, T. (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc Natl Acad Sci U S A,* 75**,** 2868-72.

Kimura, M. & Weiss, G.H. (1964) The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics,* 49**,** 561-76.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Wong, W.S., Sigurdsson, G., Walters, G.B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D.F., Helgason, A., Magnusson, O.T., Thorsteinsdottir, U. & Stefansson, K. (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature,* 488**,** 471-5.

Kropelin, S., Verschuren, D., Lezine, A.M., Eggermont, H., Cocquyt, C., Francus, P., Cazet, J.P., Fagot, M., Rumes, B., Russell, J.M., Darius, F., Conley, D.J., Schuster, M., Von Suchodoletz, H. & Engstrom, D.R. (2008) Climate-driven ecosystem succession in the Sahara: the past 6000 years. *Science,* 320**,** 765-8.

Lacau, H., Bukhari, A., Gayden, T., La Salvia, J., Regueiro, M., Stojkovic, O. & Herrera, R.J. (2011) Y-STR profiling in two Afghanistan populations. *Leg Med (Tokyo),* 13**,** 103-8.

Lachance, J. & Tishkoff, S.A. (2013) SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Bioessays,* 35**,** 780-6.

Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D. (2012) Inference of population structure using dense haplotype data. *PLoS Genet,* 8**,** e1002453.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. & Myers, R.M. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science,* 319**,** 1100-4.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., Taha, A., Shaari, N.K., Raja, J.M., Ismail, P., Zainuddin, Z., Goodwin, W., Bulbeck, D., Bandelt, H.J., Oppenheimer, S., Torroni, A. & Richards, M. (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science,* 308**,** 1034-6.

Mcdougall, I., Brown, F.H. & Fleagle, J.G. (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature,* 433**,** 733-6.

Mele, M., Javed, A., Pybus, M., Zalloua, P., Haber, M., Comas, D., Netea, M.G., Balanovsky, O., Balanovska, E., Jin, L., Yang, Y., Pitchappan, R.M., Arunkumar, G., Parida, L., Calafell, F. & Bertranpetit, J. (2012) Recombination gives a new insight in the effective population size and the history of the old world human populations. *Mol Biol Evol,* 29**,** 25-30.

Mendez, F.L., Krahn, T., Schrack, B., Krahn, A.M., Veeramah, K.R., Woerner, A.E., Fomine, F.L., Bradman, N., Thomas, M.G., Karafet, T.M. & Hammer, M.F. (2013) An African

American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am J Hum Genet,* 92**,** 454-9.

Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L. & Reich, D. (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet,* 7**,** e1001373.

Nei, M. (1972) Genetic distance between populations. *The American Naturalist,* 106**,** 283-292.

Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A,* 70**,** 3321-3.

Nei, M. & Roychoudhury, A.K. (1974) Sampling variances of heterozygosity and genetic distance. *Genetics,* 76**,** 379-90.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M. & Bustamante, C.D. (2008) Genes mirror geography within Europe. *Nature,* 456**,** 98-101.

Novembre, J. & Ramachandran, S. (2011) Perspectives on human population structure at the cusp of the sequencing era. *Annu Rev Genomics Hum Genet,* 12**,** 245-74.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. & Reich, D. (2012) Ancient admixture in human history. *Genetics,* 192**,** 1065-93.

Pickrell, J.K. & Pritchard, J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet,* 8**,** e1002967.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'connor, T.D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager,

A.E., Malig, M., Hernandez-Rodriguez, J., Hernando-Herraez, I., Prufer, K., Pybus, M., Johnstone, L., Lachmann, M., Alkan, C., Twigg, D., Petit, N., Baker, C., Hormozdiari, F., Fernandez-Callejo, M., Dabad, M., Wilson, M.L., Stevison, L., Camprubi, C., Carvalho, T., Ruiz-Herrera, A., Vives, L., Mele, M., Abello, T., Kondova, I., Bontrop, R.E., Pusey, A., Lankester, F., Kiyang, J.A., Bergl, R.A., Lonsdorf, E., Myers, S., Ventura, M., Gagneux, P., Comas, D., Siegismund, H., Blanc, J., Agueda-Calpena, L., Gut, M., Fulton, L., Tishkoff, S.A., Mullikin, J.C., Wilson, R.K., Gut, I.G., Gonder, M.K., Ryder, O.A., Hahn, B.H., Navarro, A., Akey, J.M., Bertranpetit, J., Reich, D., Mailund, T., Schierup, M.H., Hvilsom, C., Andres, A.M., Wall, J.D., Bustamante, C.D., Hammer, M.F., Eichler, E.E. & Marques-Bonet, T. (2013) Great ape genetic diversity and population history. *Nature,* 499**,** 471-5.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-59.

Quintana-Murci, L., Semino, O., Bandelt, H.J., Passarino, G., Mcelreavey, K. & Santachiara-Benerecetti, A.S. (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet,* 23**,** 437-41.

Razeto-Barry, P., Diaz, J. & Vasquez, R.A. (2012) The nearly neutral and selection theories of molecular evolution under the fisher geometrical framework: substitution rate, population size, and complexity. *Genetics,* 191**,** 523-34.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.J., Kelso, J., Slatkin, M. & Paabo, S. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature,* 468**,** 1053-60.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. & Lander, E.S. (2001) Linkage disequilibrium in the human genome. *Nature,* 411**,** 199-204.

Rosenthal, P.J. (2011) Lessons from sickle cell disease in the treatment and control of malaria. *N Engl J Med,* 364**,** 2549-51.

Schwarcz, H.P. & Grun, R. (1992) Electron spin resonance (ESR) dating of the origin of modern man. *Philos Trans R Soc Lond B Biol Sci,* 337**,** 145-8.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C. & Froguel, P. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature,* 445**,** 881-5.

Slatkin, M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics,* 139**,** 457-62.

Smith, T.M., Tafforeau, P., Reid, D.J., Grun, R., Eggins, S., Boutakiout, M. & Hublin, J.J. (2007) Earliest evidence of modern human life history in North African early Homo sapiens. *Proc Natl Acad Sci U S A,* 104**,** 6128-33.

Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S., Baker, A., Snorradottir, S., Bjarnason, H., Ng, M.C., Hansen, T., Bagger, Y., Wilensky, R.L., Reilly, M.P., Adeyemo, A., Chen, Y., Zhou, J., Gudnason, V., Chen, G., Huang, H., Lashley, K., Doumatey, A., So, W.Y., Ma, R.C., Andersen, G., Borch-Johnsen, K., Jorgensen, T., Van Vliet-Ostaptchouk, J.V., Hofker, M.H., Wijmenga, C., Christiansen, C., Rader, D.J., Rotimi, C., Gurney, M., Chan, J.C., Pedersen, O., Sigurdsson, G., Gulcher, J.R.,

Thorsteinsdottir, U., Kong, A. & Stefansson, K. (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet,* 39**,** 770-5.

Stringer, C. (2003) Human evolution: Out of Ethiopia. *Nature,* 423**,** 692-3, 695.

Stringer, C. (2012) Evolution: What makes a modern human. *Nature,* 485**,** 33-5.

Sved, J.A., Mcrae, A.F. & Visscher, P.M. (2008) Divergence between human populations estimated from linkage disequilibrium. *Am J Hum Genet,* 83**,** 737-43.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-73.

Thorne, A.G. & Wolpoff, M.H. (1992) The multiregional evolution of humans. *Sci Am,* 266**,** 76-9, 82-3.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A.C. (1991) African populations and the evolution of human mitochondrial DNA. *Science,* 253**,** 1503-7.

Weaver, T.D. (2012) Did a discrete event 200,000-100,000 years ago produce modern humans? *J Hum Evol,* 63**,** 121-6.

White, T.D., Asfaw, B., Degusta, D., Gilbert, H., Richards, G.D., Suwa, G. & Howell, F.C. (2003) Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature,* 423**,** 742-7.

Wilson, I., Weale, M. & Balding, D. (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Statist. Soc. A,* 166**,** 155-201.

Wright, S. (1931) Evolution in Mendelian Populations. *Genetics,* 16**,** 97-159.

Wright, S. (1943) Isolation by Distance. *Genetics,* 28**,** 114-38.

Wright, S. (1951) The genetical structure of populations. *Annals Eugenics,* 15**,** 323-354.

Xing, J., Watkins, W.S., Shlien, A., Walker, E., Huff, C.D., Witherspoon, D.J., Zhang, Y., Simonson, T.S., Weiss, R.B., Schiffman, J.D., Malkin, D., Woodward, S.R. & Jorde, L.B. (2010) Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics,* 96**,** 199-210.

Zalloua, P.A., Xue, Y., Khalife, J., Makhoul, N., Debiane, L., Platt, D.E., Royyuru, A.K., Herrera, R.J., Hernanz, D.F., Blue-Smith, J., Wells, R.S., Comas, D., Bertranpetit, J. & Tyler-Smith, C. (2008) Y-chromosomal diversity in Lebanon is structured by recent historical events. *Am J Hum Genet,* 82**,** 873-82.

Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R., Rayner, N.W., Freathy, R.M., Barrett, J.C., Shields, B., Morris, A.P., Ellard, S., Groves, C.J., Harries, L.W., Marchini, J.L., Owen, K.R., Knight, B., Cardon, L.R., Walker, M., Hitman, G.A., Morris, A.D., Doney, A.S., Mccarthy, M.I. & Hattersley, A.T. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science,* 316**,** 1336-41.