Demography and genetic adaptation: examples from human populations

Isabel Mendizabal Eceizabarrena

TESI DOCTORAL UPF / 2012

THESIS DIRECTOR

Dr. David Comas

Ciències Experimentals i de la Salut

UNIVERSITAT POMPEU FABRA

Urkori
Atta ta Amari

Acknowledgments

Volia agrair primer a tot Bioevo en conjunt per acollir-me tant bé durant aquest temps. Han estat uns anys inoblidables per a mi. Em sento una gran privilegiada d'haver conegut tanta gent de qui he pogut aprendre tant. Gràcies de veritat, us trobaré molt a faltar.

Moltes gràcies especialment al David, per aquests anys ("*sei*", que es diu aviat), per confiar en mi i ser el meu mentor incondicional. Gràcies per la disponibilitat, la confiança, el suport i l'apreci que m'has demostrat tant a la feina com a nivell personal.

Gracias a Óscar Lao, por tu generosidad, calidad científica y humana.

Al Francesc, per la teva passió i dedicació per la ciència. A l'Arcadi per fer que "Evolució" fos la millor assignatura de la carrera i qui va motivar que anès a Bioevo. A en Jaume per ser un bon "patriarca". Gràcies a tots aquells amb qui he tingut feedback científic, i als qui han participat activament als book clubs, journal clubs i seminaris.

Gràcies Mònica, per cuidar-me durant tots aquests anys. Karla, por tu amistad. Judit, por ser la más chungui. Laura, per haver-te tingut al meu costat. Bego, por todas nuestras discusiones. Ludovica y a Bruno, per ser la millor companyia. Sisters Belentxu y Marta, por todas las tonterías (y vinos) que nos han hecho reír tanto. Óscar, per ser tan entranyable. Txemita, per ser tan autèntic. Gracias a las flower-power Elena y Paula por hacer crecer

flores en el cemento. Ferran i Roger per ser tan 'ximples'. A Pierre, María, Fede y Johannes por vuestro buen rollo. A Graciela por tu bondad. A Rui, por tus inesperadas visitas. A las italianas inolvidables: Valeria, Chiara y Antonella. Gràcies a l'Abril i Diego pel vostre carinyo. A l'Oriol per ser únic i irrepetible. Thanks to the new generation Marc, Michael and Arturo. Gràcies als bioevos que han marxat (David Soria 'debuti!', Martin, Michelle, Gemma, Araceli, Stéphie, Andrés, Anna, Ville etc) i a tots els que em deixo (que sou molt(e)s!). Us espero allà on sigui.

A los futboleros de todos estos años, desde nuestros inicios más patatas (con piscinazos para la posteridad ¡eh Judit!), hasta el nivel super-profesional actual. A las doñas del paddle. Als bioevoleiros debutants (amb el estil "que quema que quema" i els crits de 'fuig, fuig!' d'en Joan), als pitufos, naranjitos i beachpumbas.

Per tots els "Crazy Friday"s, moments Bitácora, Filferro, Absenta, 'Gallegues', Can Maño, mojitos 2x1, Cangrejo & Arena, kebabs, viatges, i festes.

Gràcies als companys de la uni.

Thanks to all the people in Manfred Kayser's lab in the Erasmus MC in Rotterdam for receiving me so well, especially to Óscar, Andreas, and Mannis. Gracias a Lourdes y Óscar por abrirme las puertas de vuestra casa.

To Karima and Cristina, for the short but good time we worked together.

Eskerrik asko bihotz-bihotzez,

Etxekoei. Iñigori, nere lanarekiko beti erakutsi izan dezun interesagatik. Amari, nere erabakiak beti onartu izan dituzulako. Attari, nere erabakietan gogor egiteagatik. Eskerrak famili guztiari eta Izeba Marijori bereziki, zure aholkuengatik. Egunen batean zuengandik urruti egon izanak mereziko duelakoan.

Kuadrillari. Zuekin egoteko neuzkan ordu gehiegi lanari eskaini dizkiotelako, tesi honen zatitxo bat zuena da.

Neka, Eider eta Uxueri, "Siziliako" urte zoragarri haiengatik, eta tesi hau hasteko emandako indarragatik.

Aratz, Gerard, Uxue eta Volkerri, Bartzelonako kuadrillari, zuen aholkuengatik eta hainbeste momentutan zentzutasuna emateagatik. Azken urte hauetan nere bizitzan oso garrantzitsuak izan zaretelako. Faltan botako zaituztet.

Urkori bereziki. Eskerrik asko nerekin zaitudalako eta eguneroko bizipozagatik. Tesi hau zu gabe ezinezkoa izango litzatekeelako, hainbestetan galdutako indarra emateagatik eta egunero zugandik ikaragarri ikasten dudalako. Aukeratu dugun bidai honek nora garamatzan jakin ez arren zurekin edonora joango nintzatekeelako.

During this work I was supported by a PhD fellowship from the Basque Government (Eusko Jaurlaritza, Hezkuntza, Hezkuntza, Unibertsitate eta Ikerketa Saila, ikertzaileak prestatzeko laguntza BFI07.4).

Abstract

The human colonization of worldwide landmasses occurred through complex patterns of dispersal and admixture. At the same time, the survival in the different areas of the world depended on the adaptation to new habitats that imposed novel selective challenges. With the advent of high-throughput genotyping technologies and dense catalogues of human genetic variation, the demographic history of many human populations has been unraveled from genomic data, with important implications in medical genetics. However, several human groups are yet to be genetically characterized. These incomplete past histories include the determination of ancestries of the current Cuban population, as well as the origins and dispersal of European Romani, whose demographic history is aimed to be reconstructed in this work. Finally, the present study also aims to describe the genetic basis and evolution of one of the most striking human phenotypes, the African Pygmy height.

Resumen

La colonización humana de las diferentes masas continentales se produjo mediante complejos patrones de dispersión y mezcla. La supervivencia en las diversas regiones del planeta ha dependido de la adaptación a las presiones selectivas impuestas por los nuevos hábitats. Con el desarrollo de tecnologías de genotipado masivo y las bases de datos de la diversidad genética humana, la historia demográfica de muchas poblaciones humanas, y sus implicaciones médicas, han sido descritas. Sin embargo, algunas poblaciones todavía no han sido caracterizadas genéticamente. Por ejemplo, tanto la descomposición de la ancestría genética de la población cubana actual como los orígenes y la dispersión de los gitanos europeos siguen siendo historias incompletas que se han reconstruido en esta tesis. Finalmente, este estudio también tiene como objetivo describir la evolución y las bases genéticas de uno de los fenotipos humanos más llamativos, la altura de los pigmeos africanos.

Preface

This thesis was developed in the midst of the post-genomic era. After the completion of the highly valuable reference human sequence, the burst of high-throughput genotyping technologies allowed the systematic survey of genetic variation in worldwide human populations. For the first time, the possibility to study population-specific demographic and adaptive histories at the genomic level was opened up.

The genomic inferences on human origins can yield to amazing discoveries. Recently, a strict out-of-Africa model was replaced with the exciting discovery that we carry genes from extinct hominins. Not only as a species, but also at population and individual level, the reconstruction of our origins has received important social attention, as witnessed by the increasing popularity of recreational genetic ancestry testing and scientific blogs.

Making the phenotypic interpretation of the genetic code is an important requirement for fulfilling the promises of medical genetics. However, in face of the bulk of genomic data, scientists realized about the poor functional understanding of the genome. Making the link between genomes and phenotypes with current technologies is not straightforward. Population genetics field opens the exciting possibility to scan the whole genome to find molecular footprints of selective events and shed light into the mechanistic basis of adaptation. Indirectly, this approach can provide very valuable phenotypic information and point at genomic basis of important evolutionary traits.

By producing new valuable data and taking profit of that accumulated so far, this work aims to reconstruct interesting instances of human population histories such as those of Cuban and European Romani, as well as to find the genetic basis of the African Pygmy height while exploring its adaptive value.

Index

	Page
Abstract	ix
Resumen	xi
Preface	xiii
1. Introduction	1
1.1. Forces shaping genetic variation	
1.1.1. Mutation	3
1.1.2. Recombination	6
1.1.3. Genetic drift	8
1.1.4. Migration	10
1.1.5. Natural selection	10
1.1.6. Interplay between evolutionary forces	11
1.2. Inferring demographic history from genetic variation	
1.2.1. The coalescent theory	13
1.2.2. Analyzing population structure	15
1.2.3. Bayesian inference of evolutionary parameters	20
1.2.4. Phylogeography: mtDNA and Y-chromosome	23
1.3. Inferring adaptation from genomic variation	
1.3.1. Methods for detecting classical selective sweeps	27
1.3.2. Adaptation beyond classical sweeps	31
1.4. Human genetic variation	
1.4.1. Origin of Anatomically Modern Humans	33
1.4.2. The peopling of Cuba	35
1.4.3. Demographic history of European Romani (Roma)	36
1.4.4. Demographic and adaptive history of African Pygmies	39
1.4.5. Public resources of human DNA variation	42

2. Objectives 47

3. Results 51

3.1. Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba 53

3.2. Reconstructing the Indian origin and dispersal of the European Roma: A maternal genetic perspective 65

3.3. Reconstructing the population history of European Romani from genome-wide data 77

3.4. Adaptive evolution of loci covarying with the human African Pygmy phenotype 97

4. Discussion 113

Bibliography 137

Appendix. Contributions to other articles

A1. SNPlexing the human Y-chromosome: A single-assay system for major haplogroup screening 155

A2. Genetic structure of Tunisian ethnic groups revealed by paternal lineages 163

A3. Y-chromosome diversity in Native Mexicans reveals continental transition of genetic structure in the Americas 175

Electronic Appendix. Supplementary information for results

B1. Supplementary information for section 3.1 Attached CD-ROM

B2. Supplementary information for section 3.2

B3. Supplementary information for section 3.3

B4. Supplementary information for section 3.4

1. Introduction

1.1. Forces shaping genetic variation

The goal of population genetics is to understand the forces that generate and maintain genetic diversity within species. Understanding the mechanisms of action of these forces enables the building of models to explain observed diversity patterns and to infer past processes. The ultimate force that generates variation is mutation; recombination creates new allele combinations, whereas demography (genetic drift and migration) and natural selection change the allele frequencies through time. In this section, these evolutionary forces will be briefly described.

1.1.1. Mutation

Mutation (any permanent change in the DNA molecule) is the force that provides the raw material on which other evolutionary forces act. From an evolutionary point of view, only genomic changes that are heritable and pass through the germ-line are of interest. Ranging from single base changes to large chromosomal events, different types of mutations can be defined depending on the type of structural change in the DNA (see Figure 1 for an overview). At the population level, a variant found above 1% of frequency is called "polymorphic", "common" if its prevalence is higher than 5% and "rare" otherwise.

a) Point mutations

Point mutations or single nucleotide polymorphisms (SNP) are substitutions of a single base, in which one nucleotide is exchanged for another. SNPs are the most common type of mutations and are typically binary markers, i.e. biallelic. In average, two humans differ in 1 in 1,000 nucleotides and, at the time of writing, there were 41 million known SNPs in humans (dbSNP build 135). Nevertheless, this number is expected to rise in the next few years with the ongoing wave of resequencing studies.

Base substitutions from one pyrimidine (C and T) to another pyrimidine or from a purine (A and G) to a purine are called "transition". Exchanges from pyrimidine to a purine or vice versa are called "transversion". Within a population, the allele with the highest prevalence is called "major allele" whereas the other is considered the "minor allele" and its frequency is known as MAF ("Minor Allele Frequency"). In addition, if the ancestral state of the position is known, the young allele is called "derived" in opposition to the "ancestral" allele. Finally, several types of SNPs can be distinguished regarding the functional implications of the point mutations in the open reading frame. When a SNP happens in a region that does not codify for proteins is called "non-coding". A "coding" SNP can be "non-synonymous" if the mutation alters the amino acid ("missense" if it changes the amino acid, "nonsense" if it truncates the protein, "frameshift" if it introduces a change in the reading frame, see Figure 1) or "synonymous" if the mutation is silent and the protein sequence remains unchanged due to the redundancy of the genetic code.

Since genomic mutation rates are usually low ($\mu=1.2x10^{-8}$ per generation (The 1000 Genomes Project Consortium 2011)), it is unlikely that any mutation at a given position have recurred over the time-scale of the evolution of modern humans. Consequently, and despite some known exceptions (i.e. the HVS-I in the control region of mitochondrial DNA, with $\mu = 4x10^{-6}$ per generation (Soares et al. 2009)), it can be assumed that point mutations show identity by descent (IBD) rather than identity by state (IBS). That is, if two chromosomes show the same allele it is more likely that they have inherited it from a common ancestor. This is the reason why SNPs are sometimes referred to as unique event polymorphisms (UEPs). Thus, the infinite site model is suitable for point mutations since it assumes that the genome is infinitely long so that each mutation occurs in a previously non mutated site (Kimura 1969).

b) Insertions and deletions

Currently, the term "indel" is applied to insertion-deletion variations in DNA that are smaller than 50 base pairs (bp) in length (Alkan et al. 2011). One type of indels are short tandem repeats (STR or microsatellites), which are arrays of consecutive repeats of 1-6 bp units. STRs are ubiquitous in eukaryotes, and especially in mammal genomes (Levinson and Gutman 1987). Human Y-chromosome STRs are widely used in forensics, paternity tests, and genealogical DNA testing. Specifically, they take profit of the typically high mutation rates of indels (μ=6.9x10^{-4} per generation (Zhivotovsky et al. 2004)) that results in highly diverse patterns in evolutionary short time periods. For STR data, stepwise mutation model (SMM) (Di Rienzo et al. 1994; Ohta and Kimura 1973) is more appropriate, as it considers that the allele length increases or decreases in a number of bp in each mutation.

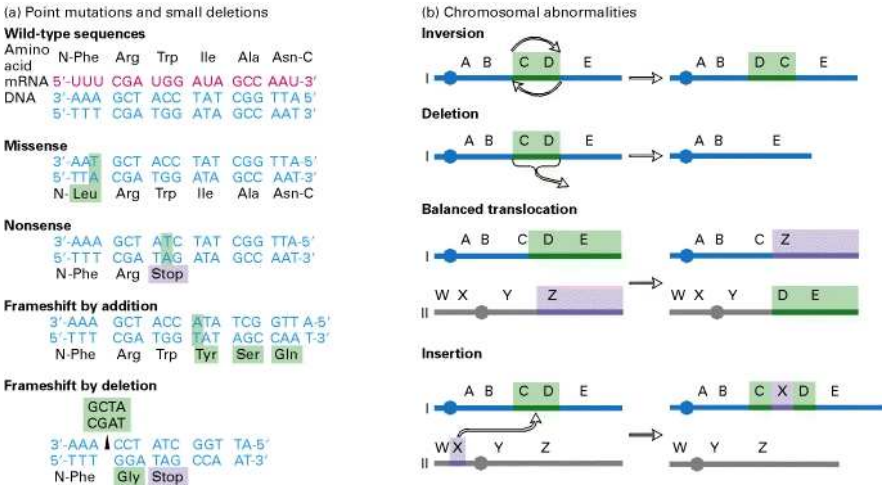


Figure 1. Types of mutations in the DNA. Altered nucleotides and amino acids are highlighted in green and purple. From Lodish et al. (2000).

c) Structural variation

Structural variants are genomic rearrangements larger than 50 bp (Alkan et al. 2011). These include deletions, novel insertions, mobile-element

transpositions, duplications, and translocations. The advent of genome-scanning technologies showed that structural variants were unexpectedly abundant in the human genome, revealing that the human reference sequence was incomplete regarding structural variation (Kidd et al. 2008). In the past years we have witnessed an unprecedented increase of orders of magnitude in the discovery of structural variants (Alkan et al. 2011). Despite not being covered in this work, advances in the understanding of structural variation in the near future are likely to make important contributions in the study of human adaptation and disease susceptibility (Iskow et al. 2012).

1.1.2. Recombination

Most of our genome is biparentally inherited (except the mitochondrial DNA and the non-recombinant parts of the Y and X chromosomes) and undergoes the process of recombination. Recombination is the mechanism by which maternal and paternal homologous chromosomes align and exchange segments during meiosis. Therefore, different alleles are reshuffled making new combinations in the same DNA molecule, named haplotype. A basic consequence of recombination is linkage disequilibrium (LD). That is, some haplotypes may be more frequent than expected from the relative frequencies of the alleles. Recombination rates are not uniform across the genome and the presence of recombination hotspots shapes the genome in a block-like manner, with blocks showing high internal LD separated by other blocks by low LD between them (Reich et al. 2001).

Nevertheless, LD not only depends on recombination, but also on the mutational history as outlined in Figure 2. Since LD is responsible for genomically linked loci to share a common evolutionary history, the evolutionary forces that shape variation, such as drift and selection, can also leave a footprint on LD patterns. Finally, gene flow between two populations

presenting differences in allele frequencies also creates LD after the admixture event (Pfaff et al. 2001).

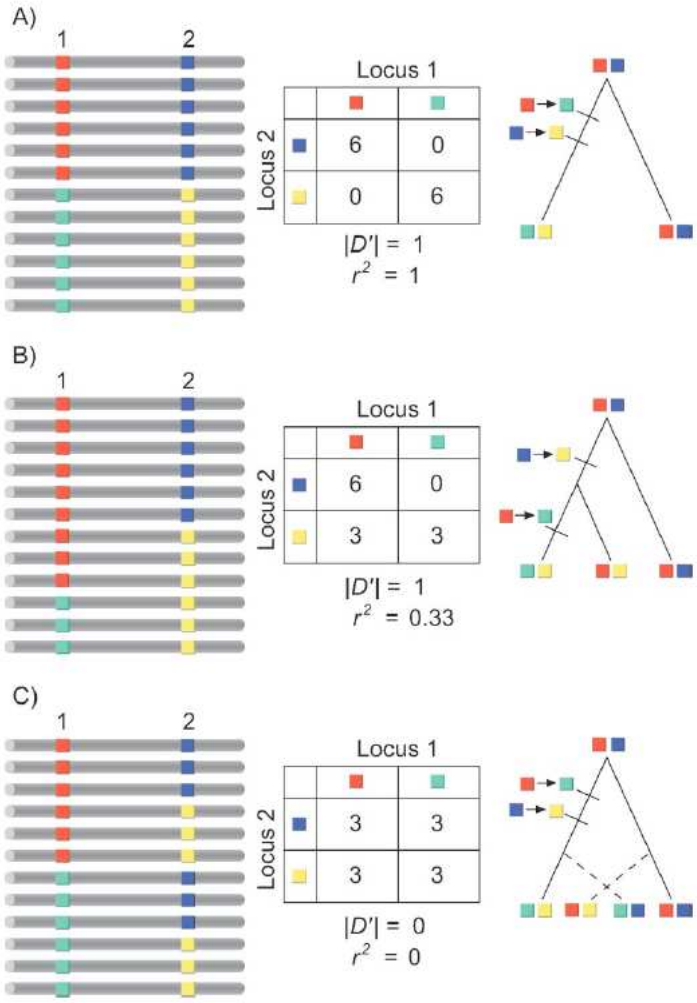


Figure 2. Scenarios of linkage disequilibrium (LD). The figure shows three different cases of LD and the behaviour of D' and r^2 statistics (widely used LD measures). Images in the left column represent the allelic states of two loci. The middle column represents the 2x2 contingency table of haplotypes and the resulting r^2 and *D'* statistics. The right column represents a possible tree responsible for the observed LD present. **A)** Absolute LD exists when two loci share a similar mutational history with no recombination. Both r^2 and *D*' equal 1. **B)** LD can result when mutations occur on different lineages without recombination between the loci. Notice the large difference in measures of LD as calculated by r^2 and *D*'. **C)** Linkage equilibrium is produced when there is recombination between loci, regardless of mutational history. In this situation, both r^2 and *D*' equal 0. From Flint-Garcia et al. (2003).

1.1.3. Genetic drift

In an imaginary population of infinite size without mutation, migration or natural selection, allele frequencies remain the same from one generation to another. However, real populations are of finite size and each generation represents a sampling from the previous one, introducing stochastic variation. Genetic drift (Wright 1931) can be defined as the joint effect of randomness in the birth and death of individuals in a population. This includes the sampling of gametes from one generation to another, the variation in the number of offspring by carriers of different alleles or the variation in the number of those that survive to reproduce. This stochastic force makes allele frequencies fluctuate in time until alleles are ultimately lost or fixed.

The basic model for reproduction in a finite population is the Wright-Fisher model (Fisher 1930; Wright 1931). The populations are assumed of constant size (N), with random mating and non-overlapping generations. In this model, the N individuals that will form the following generation are obtained by sampling 2N independent gametes from the current population by binomial sampling. The only factor that affects the allele frequencies in the future generation is the current allele frequency. In spite of the unrealistic assumptions and the randomness of the process, the Wright-Fisher model provides a complete description of how allele frequencies drift in populations with time (Figure 3). Genetic drift is closely related to population structure as, after divergence, populations fix different alleles by chance. Indeed, natural populations typically show certain degree of isolation (limited gene flow among subpopulations) and consequently, members of the same subpopulation are more closely related on average than those from different subpopulations.

The model also predicts that the magnitude of genetic drift (the speed of allele loss or fixation) is directly related with the population size (N) that is being sampled. The smaller the population, the higher the probability of new mutations to reach fixation (equal to its frequency = 1/2N). A more useful measure of magnitude of drift effects is provided by the effective populations size ($N_{e,}$) (Wright 1931). N_e represents the size of a Wright-Fisher population that experiences the same amount of genetic drift as the one observed in the population of interest. Therefore, N_e allows the comparison of drift effects experienced by different populations (independently of current population sizes) or different loci in the genome (such as mtDNA, with a fraction of ¼ transmitted copies than that of autosomes). The average time of fixation can be calculated as $t = 4N_e$ generations (Kimura and Ohta 1969). Populations that undergo demographic events such as founder episodes and population bottlenecks experience strong drift processes, and exhibit reduced levels of genetic diversity.

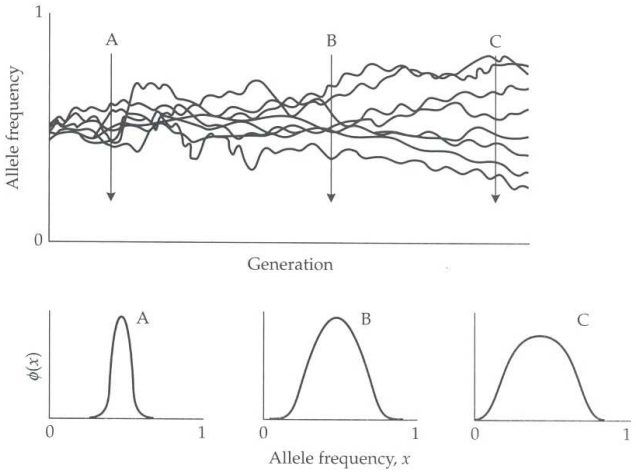


Figure 3. Genetic drift and population structure. The figure illustrates the implications of random genetic drift in different subpopulations undergoing the process of repeated sampling. The top figure shows how allele frequencies in each subpopulation change erratically, making them drift apart with time. Bottom figure shows that the variance of the distribution of allele frequencies in the subpopulations increases over time. From Hartl and Clark (1980).

1.1.4. Migration

Migration is the movement of individuals among subpopulations. This results in gene flow between them, decreasing their genetic differentiation. Several mathematical models have been proposed for migration. The island model (Wright 1931) assumes a meta-population splits into different equal sized demes with equal migration rates; the stepping-stone model (Kimura and Weiss 1964) introduces the idea of geographic structure by allowing gene flow only between geographically adjacent demes. Finally, in the isolation by distance model (Wright 1943), genetic similarity is a function of the geographical distance between the demes. Nevertheless, these models are necessarily simpler than actual human migration patterns, where migrants may not be a random sampling of the subpopulation as regards to sex, age or kinship (Jobling et al. 2004).

1.1.5. Natural selection

Natural (Darwinian) selection can be defined as the heritable variation in reproductive success. That is, different genotypes provide different capacities to survive and reproduce in a certain environment (fitness). Adaptation is the movement of a population towards a phenotype that best fits the current environment (Fisher 1930). Importantly, selection acts on the phenotype and not on the genotype, which apart from genetic interactions may also involve environmental factors.

The relative fitness of the fittest genotype compared to the competing others is called selection coefficient (s). Different types of selection can be defined depending if the fitness of the carrier of the genotype is reduced by the effect of selection (negative) or increased (positive). In those scenarios, the allele frequencies of the selected alleles will tend to decrease or increase respectively during generations. Balancing selection is a type of selection that favours balanced polymorphism in time, increasing variability in the

population. This includes cases of overdominant selection (when the heterozygote is the fittest genotype) and frequency-dependent selection (highest fitness when a genotype is at low frequencies). Even weak selective forces can cause appreciable changes in allele frequencies over generations.

1.1.6. Interplay between evolutionary forces

"Survival of the fittest", Herbert Spencer , 1864
"Survival of the luckiest", Motoo Kimura, 1989

As molecular data started to accumulate in the 60s, the observed amount of polymorphism exceeded the expected value under the strong selectionist view dominant at that time. Kimura posed an explanation with the neutral theory of molecular evolution (1968), which states that the vast majority of the observed genomic variation is neutral (no fitness effect) and governed by genetic drift. The model allows a role of negative selection to eliminate the deleterious mutations that may appear in functional regions, but considers that positive selection events are rare.

According to this theory, the amount of expected polymorphism can be predicted from the interplay between mutation and drift. The expected diversity found in a diploid population that is in mutation-drift equilibrium (in balance between newly generated alleles by mutation at rate μ and those lost by drift) is the population mutation parameter theta, $\theta = 4N_e\mu$. Theta can be estimated by different methods based on different aspects of genetic diversity such as the number of alleles, the number of segregating sites (S), the observed homozygosity, and the number of pairwise differences (π). While these measures should give the same value under neutrality, different demographic and selective factors may alter the expectations of these measures. Thus, the neutral theory provides the theoretical background for

modelling the effects of demography (next section) and for detecting departures of the model to infer selection (section 1.3).

1.2. Inferring demographic history from variation

As shown in the previous section, population genetics provides the theoretical background on how different processes shape genomic diversity. In the current post-genomic time, models such as the coalescent and new statistical methods have been successfully applied to infer past demographic events from genomic data.

1.2.1. The coalescent theory

The coalescent theory (Hudson 1990; Kingman 1982) provides a mathematical description of the genetic ancestry of a DNA sample (gene genealogy) as we move backwards in time. The sample is assumed to evolve according to a neutral Wright–Fisher model. Going back in time two random lineages can merge in the same ancestor (coalescence event). The process continues until the whole sample coalesces in a single common ancestor, which is an efficient approximation to the ancestor of the entire population (Figure 4a-c). There are two key parameters in this process. The first is the rate of coalescence, which is inversely proportional to the population size, and the second is mutation rate. Since variation does not affect fitness, the structure of the tree and mutation can be treated separately in two random processes (mutations are superimposed forwards on the generated tree). This feature makes the process much more efficient from a computational point of view. The pattern of polymorphism in the sample will entirely depend on the shape of the genealogy, which is determined by genetic drift. The incorporation of recombination into the model provokes that linked fragments have different genealogies (Figure 4d).

The coalescent has become the reference model for demographic inferences in population genetics. Coalescence simulations are based on thousands of plausible gene genealogies generated stochastically under a specified

evolutionary model that are used to estimate demographic parameters or to assess the role of different evolutionary forces (see section 1.2.3).

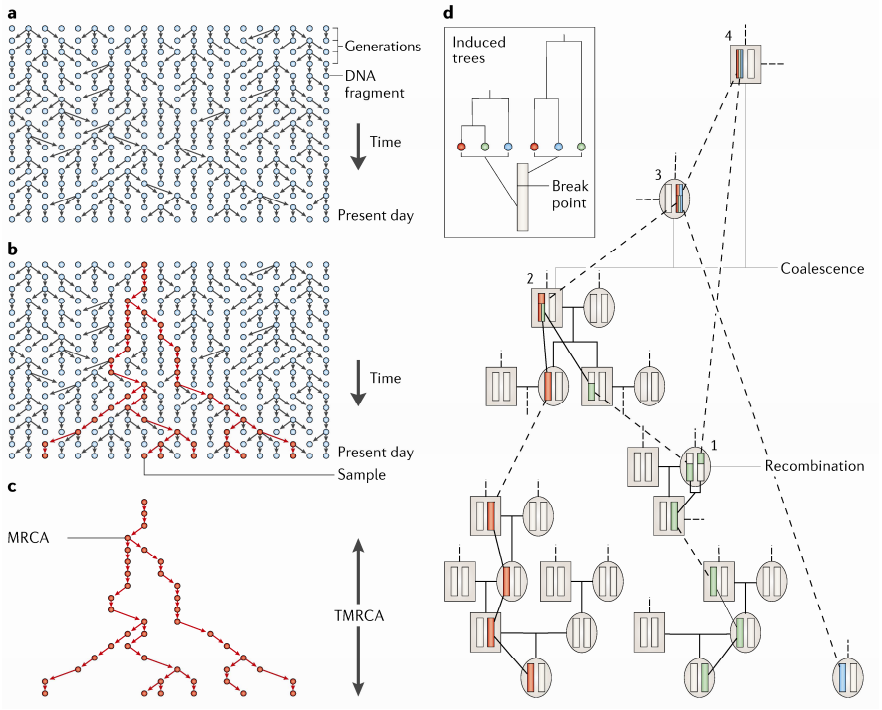


Figure 4. The coalescent. **A.** Each row represents a generation, where each blue circle is a DNA fragment. **B.** Ancestry of a sample from the present day (six fragments) is traced back in time, as indicated in red. **C.** The most recent fragment from which the entire sample is descended is known as the 'most recent common ancestor' (MRCA), and the time at which it appears is known as the 'time to the most recent common ancestor' (TMRCA). **D.** The coalescent with recombination. Lines bifurcate, as well as coalesce, as we move back in time. The genealogy for three copies of a fragment is shown. In event 1 the green lineage undergoes recombination and splits into two lineages, which are then traced separately; in event 2 one of the resulting green lineages coalesces with the red lineage, creating a segment that is partially ancestral to both green and red, and partially ancestral to red only; in event 3 the blue lineage coalesces with the lineage created by event 2, creating a segment that is partially ancestral to blue and red, and partially ancestral to all three colours; in event 4 the other part of the green lineage coalesces with the lineage created by event 3, creating a segment that is ancestral to all three colours in its entirety. As shown, the recombination event induces different genealogical trees on either side of the break. From Marjoram and Tavaré (2006).

1.2.2. Analyzing population structure

Genetic substructure can reveal genetic affinities among populations and reflect past demographic histories such as episodes of population splits, founder events, isolation, admixture, and migration.

Apportionment of genetic variation

As introduced in section 1.1.3, the division of a meta-population in partially isolated subpopulations results in differential fixation and lost of alleles. This process will cause an excess of homozygotes (and a deficiency of heterozygotes) in the meta-population compared to random mating expectations (Wright 1951). Different statistics have been proposed to measure population genetic differentiation (see Rosenberg et al. (2003)).

Among these, the inbreeding coefficient F_{ST} (Wright 1951) is a widely used measure of how genetic diversity is apportioned among different sub-populations. It is computed 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. An alternative definition of F_{ST} is the probability that two alleles sampled at random from an ancestral population are identical-by-descent. That is, the probability that two alleles share a common ancestor within the subpopulation compared to the meta-population without intervening migration or mutation. Therefore, F_{ST} can be also defined in terms of coalescence times between alleles (Slatkin 1991). Another widely used measure is the Analysis of Molecular Variance (AMOVA) (Excoffier et al. 1992). AMOVA is also applied to apportion the variance between and among sub-populations, but it can be applied to both allele frequency and molecular data.

How is genetic diversity distributed among human populations? The F_{ST} values for resequencing data between Africans and Europeans is ~0.071 (see

Table 1), which means that 7.1% of the variance in allele frequency is found between sub-populations whereas the rest 92.9% is harboured within.

	ASN	YRI
CEU	0.052	0.071
ASN	-	0.083

Table 1. F_{ST} between continental groups from resequencing data. Average F_{ST} values from ~15 million SNPs in human populations of European (CEU), African (YRI), and Asian (ASN) ancestry (The 1000 Genomes Project Consortium 2011).

These numbers clearly show a notion that was popularized by Richard Lewontin in the 70s: most human genetic diversity lies within populations and not between populations. A main conclusion of Lewontin's study (1972) was that ethnic classification was genetically meaningless, which has been lately appointed as the "Lewontin's fallacy" (Edwards 2003). In spite of being due to a small proportion of our genome, genetic data is indeed very informative to classify individuals into subpopulations. This is possible due to substructure: allele frequencies at different loci are correlated within populations. As Figure 5 illustrates, the cumulative information provided over several loci permits to assess the classification of genetic diversity in different groups.

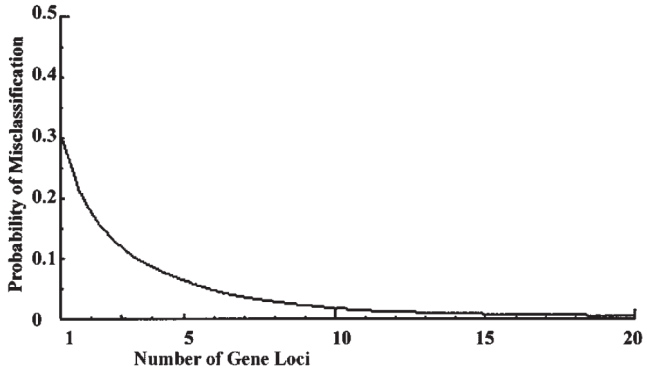


Figure 5. Probability of misclassification according to the number of loci. This example assumes two populations with frequencies of 0.3 and 0.7 for a given locus. The probability of misclassification of individuals into populations based on this locus is 0.3, and the proportion of the variability within groups is 0.84 as in Lewontin's data. However, as the number of loci increases (same frequencies) the probability of misclassification rapidly becomes negligible. From Edwards (2003).

Dimensionality reduction methods

Dimensionality reduction techniques were among the first statistical methods applied in human genetic studies to study structure (Cavalli-Sforza 1966). The objective of these methods is to summarize data with multiple variables (possibly correlated) into a few uncorrelated synthetic variables. These methods provide a simplified but more comprehensible picture of the information in the data, so that the relationships among objects (genotypes, individuals, or populations) can be revealed. In front of other methods, their advantages are that, first they are exploratory (they do not assume any specific genetic model), and second, they are reasonable in computational terms (Jombart et al. 2009).

The application of Principal component analysis (PCA, based on allele frequency data) and Multidimensional scaling plot (MDS, based on genetic distances) to study human genetic structure has become extremely popular. Typically, the objective is to represent graphically the genetic relationships between individuals on a plot so that the projected coordinates approximate the original genetic similarities (see Figure 6A). In spite of some limitations (i.e. different demographic processes can give same projections), these projections inform about the underlying genealogical history of the samples (McVean 2009). This implies that demographic inferences such as migration, geographical isolation, and admixture can be made from the study of these plots. Nevertheless, interpreting migration from gradients and wave patterns from such plots can be problematic because human genetic similarity typically decays with geographic distance (Novembre and Stephens 2008).

Clustering algorithms

Clustering methods aim to assign objects (such as individuals or populations) into clusters or groups that are more genetically similar to each other than to those in other clusters. Several clustering algorithms have been developed,

such as STRUCTURE (Falush et al. 2003; Pritchard et al. 2000) and ADMIXTURE (Alexander et al. 2009). One basic difference between these algorithms and previously shown dimensionality reduction techniques is that these are model-based, and the ancestry coefficients are parameters of a statistical model. *K* ancestral populations are assumed, each of which is characterized by a set of allele frequencies at each locus. Then, each sampled individual is assigned to a population or populations (if admixture is detected) optimizing Hardy-Weinberg and linkage equilibrium of the loci within populations. Typically, the value of *K* is unknown and several values of *K* are explored to choose that with the best fit to the data. Applications of these methods include the determination of population structure, assigning individuals to populations, and identifying migrants and admixed individuals. See Figure 6B for a typical result of these algorithms.

Another clustering method that is used in population genetics is Neighbor-Joining (Saitou and Nei 1987). In this case, the objective is to display graphically the genetic distances by drawing a bifurcating un-rooted tree (see Figure 6C). Neighbor-Joining is a computationally efficient method that aims to build the tree with the "minimum evolution" (the shortest sum of branch lengths).

The main virtue and fault of the use of clustering algorithms in population genetics is the classification of genetic variation in discrete groups. This can be very interesting in an attempt to categorize human genetic variation, but also can be lead to misinterpretation in biological terms since the actual mating patterns of human populations rarely show sharp discontinuities (see (Weiss and Long 2009) for further discussion).

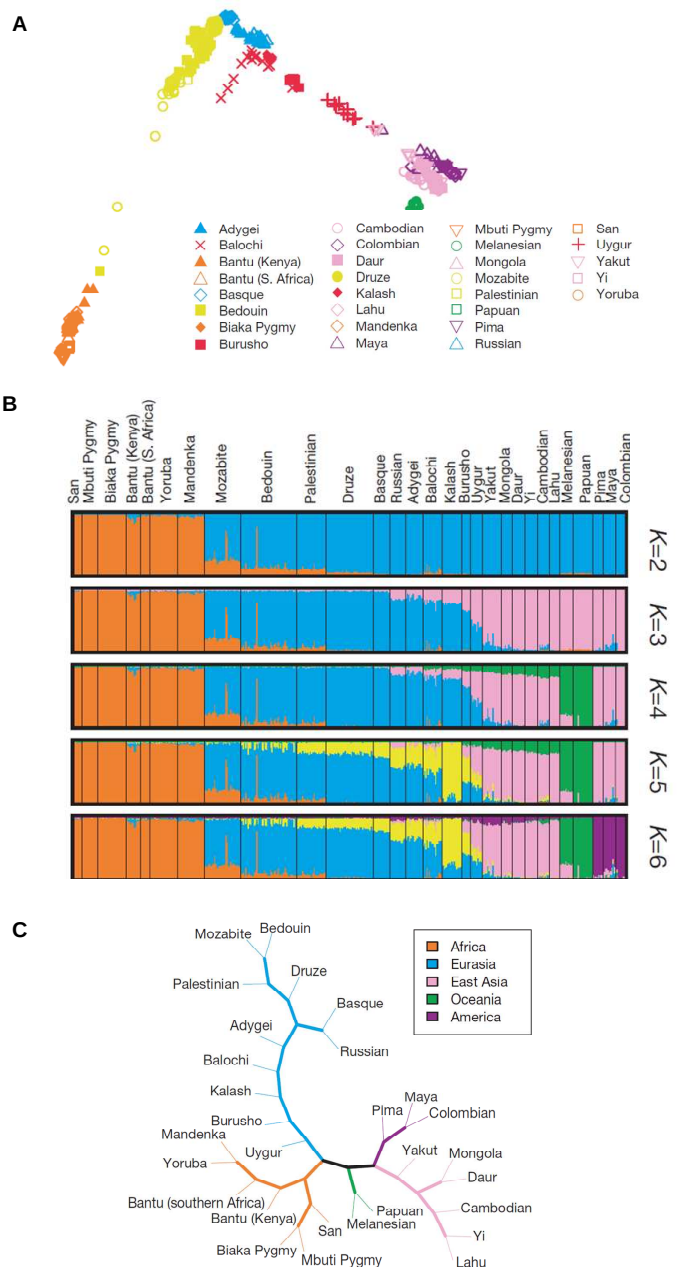


Figure 6. Examples of population structure analyses. A. MDS representation of genetic distances between individuals based on 512,762 autosomal SNPs in 443 HGDP–CEPH individuals (see section 1.4.5 for sample information). **B.** Population ancestry inferred by STRUCTURE for the same set of individuals. Each individual is shown as a thin vertical line partitioned into *K* coloured components representing inferred membership in *K* genetic clusters. **C.** Neighbour-joining tree of population relationships. Modified from Jakobsson et al. (2008).

1.2.3. Bayesian inference of evolutionary parameters

The coalescent theory has boosted the development of model-based methods that allow making inferences on how genetic data was generated. Bayesian methods incorporate prior knowledge on the parameters and update this knowledge conditioned on the observed data in the posterior distribution (see illustrative example in Figure 7).

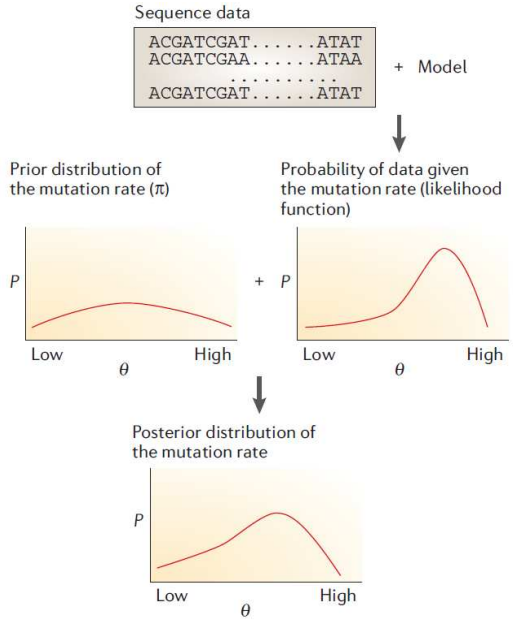


Figure 7. Model-based analysis in the Bayesian framework. The figure shows an example in which the aim is to estimate the mutation rate based on a set of mitochondrial DNA sequence data. The coalescent without recombination could be a good model to generate simulations using the mutation rates in the sequenced region (prior distribution $\pi(\theta)$). The posterior distribution for the parameter θ will be proportional to the product of the prior distribution and the likelihood (the probability of the data over the range of all possible mutation rates). From Marjoram and Tavaré (2006).

Approximate Bayesian computation

However, most models of evolution are sufficiently complex that the likelihood of the observed data given the model is unknown (Marjoram and Tavare 2006). Taking profit of the fact that simulating genetic data is

relatively easy thanks to coalescent simulators such as ms (Hudson 2002) or cosi (Schaffner et al. 2005), a solution is to generate data given a prior distribution of the parameters, and then assess how often simulated data is similar to the observed dataset. As the "approximate" word suggests, approximate Bayesian computation (ABC) methods are based on using summary statistics to evaluate the similarity between simulated and observed datasets. Thus, ABC methods allow relatively easy implementation of complex demographic scenarios using high amount of data not computationally feasible with other methods.

There are different proposed implementations of the ABC framework. The basic algorithm is the rejection algorithm (Tavare et al. 1997). The first step of the rejection algorithm consists of simulating millions of multilocus data sets. The simulations have identical sample size and loci number than the data set under study, and the parameters used are randomly sampled from prior distributions. In a second step, the simulated and observed datasets are compared by a series of summary statistics, and the similarity is quantified by means of a distance. In a third step (rejection step), those simulations with a similarity above (or distance below) a threshold to the observed data are retained, whereas the rest are discarded. The parameters that produced the ascertained simulations are a sample from the posterior distribution of the parameters given the observed data.

Other implementations of ABC based on the rejection algorithm have been proposed. For example, regression-weighted ABC (Beaumont et al. 2002) performs an additional step after the rejection. Because not all the simulations are identically close to the observed data, weighted ABC performs a weighted regression of the parameters on the distance of the summary statistics of the previously retained simulations (Figure 8).

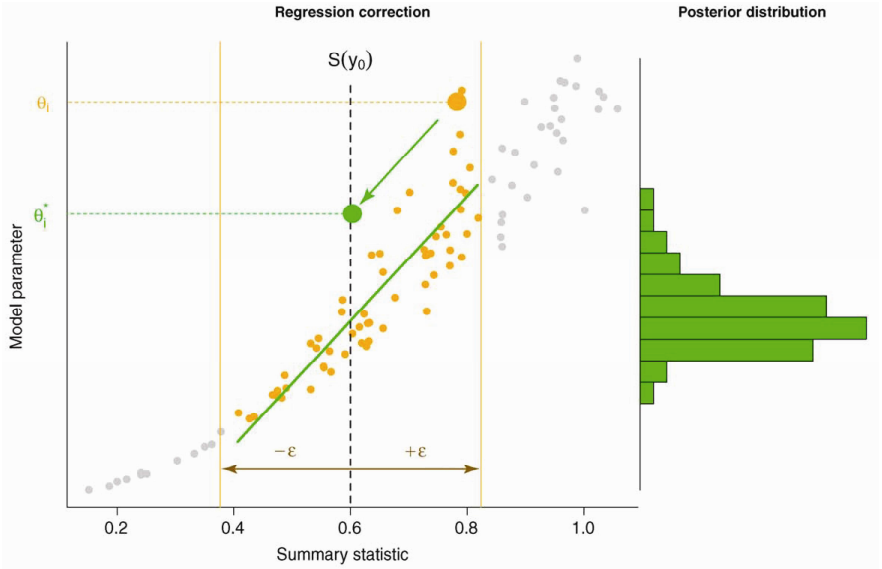


Figure 8. ABC regression-weighted algorithm. A parameter value, θ_i, is sampled repeatedly from its prior distribution to simulate a dataset, y_i, under a model. Then, from the simulated data, the value of a summary statistic, $S(y_i)$, is computed and compared with the value of the summary statistic in the observed data, $S(y_0)$, using a distance measure. If the distance between $S(y_0)$ and $S(y_i)$ is less than ε (the so-called 'tolerance'), the parameter value, θ_i, is accepted. The plot shows how the accepted values of θ_i (points in orange) are adjusted according to a linear transform, $\theta_i^* = \theta_i - b(S(y) - S(y_0))$ (green arrow), where b is the slope of the regression line. After adjustment, the new parameter values (green histogram) form a sample from the posterior distribution. From Csilléry et al. (2010).

Using the same philosophy underlying the rejection algorithm, a Markov Chain Monte Carlo (MCMC) algorithm has also been proposed for ABC (ABC-MCMC (Marjoram et al. 2003)). In this particular implementation of the ABC, the MCMC chain is started at one accepted simulation (below a certain proposed distance threshold). New parameters are proposed based on a proposal distribution. If the distance of the simulated data is below the threshold, then the chain moves to the new proposed parameters with probability proportional to its prior likelihood. Otherwise, it remains in the previous position of the parameters. By applying these steps iteratively, the posterior distributions of the parameters of interest are explored.

1.2.4. Phylogeography: mtDNA and Y-chromosome

In the last 20 years, many surveys of human demography were based on the study of non-recombinant loci. The two most widely used non-recombining systems in our genome are the mitochondrial DNA (mtDNA) and the non-recombinant part of the Y-chromosome (NRY).

The main clades (haplogroups) in the genealogical tree of the mtDNA and NRY are defined by SNPs and are assigned alphanumeric designators. Within each haplogroup, the haplotypes are related descendant mtDNA sequences or NRY STR haplotypes that evolve at higher rates allowing within haplogroup diversity comparisons. One of the most interesting characteristic of the uniparental markers is their well-resolved phylogeography (i.e. the geographic distribution of these haplogroups, see Figure 9). These trees provide a phylogeography at maximum molecular resolution, meaning that the amount of geographically informative positions that can be revealed in contiguous sequence fragments is higher than in other loci in the genome (Underhill and Kivisild 2007). Contrasting with generally low mutation rates and larger effective population sizes in the nuclear genome, the highly variable mtDNA and the NRY provide maximum information in the time window of interest for most of the human migrations around the globe (i.e. the last 200,000 years (Underhill and Kivisild 2007)).

Since they are uniparentally transmitted, these loci reflect maternal (mtDNA) and paternal (NRY) demographic histories and are therefore unique for the study of sex-biased demographic events. This interesting feature allows the study of both local cultural practices (i.e. patrilocality versus matrilocality) as well as sex differential large-scale migrations (Wilkins 2006).

However, one cannot neglect that both molecules harbour important genes, which suggests that at least negative selection (if not also positive) may have had a role on the observed diversity. This could invalidate the main assumption that present patterns of diversity were shaped by neutral forces, which is essential to draw population history inferences. Nevertheless, evidence for differential selection among different clades has been inconclusive, and the extent to which the distribution of mtDNA and NRY variation has been shaped by environmental factors remains to be assessed.

Finally, it is important to bear in mind that these two loci represent less than 2% of the genome. Due to the stochastic nature of the genealogical process, their trees may provide limited insights into the underlying demography from which they originated. To obtain a more comprehensive view of the demography history of a population, more loci might be needed. In any case, the study of mtDNA and NRY have constituted very successful to infer instances of human demographic history. Among these, studies on our African origins, continental migration routes, pre-historical expansions as well as historical migrations have been consistent with studies on multiple autosomal loci.

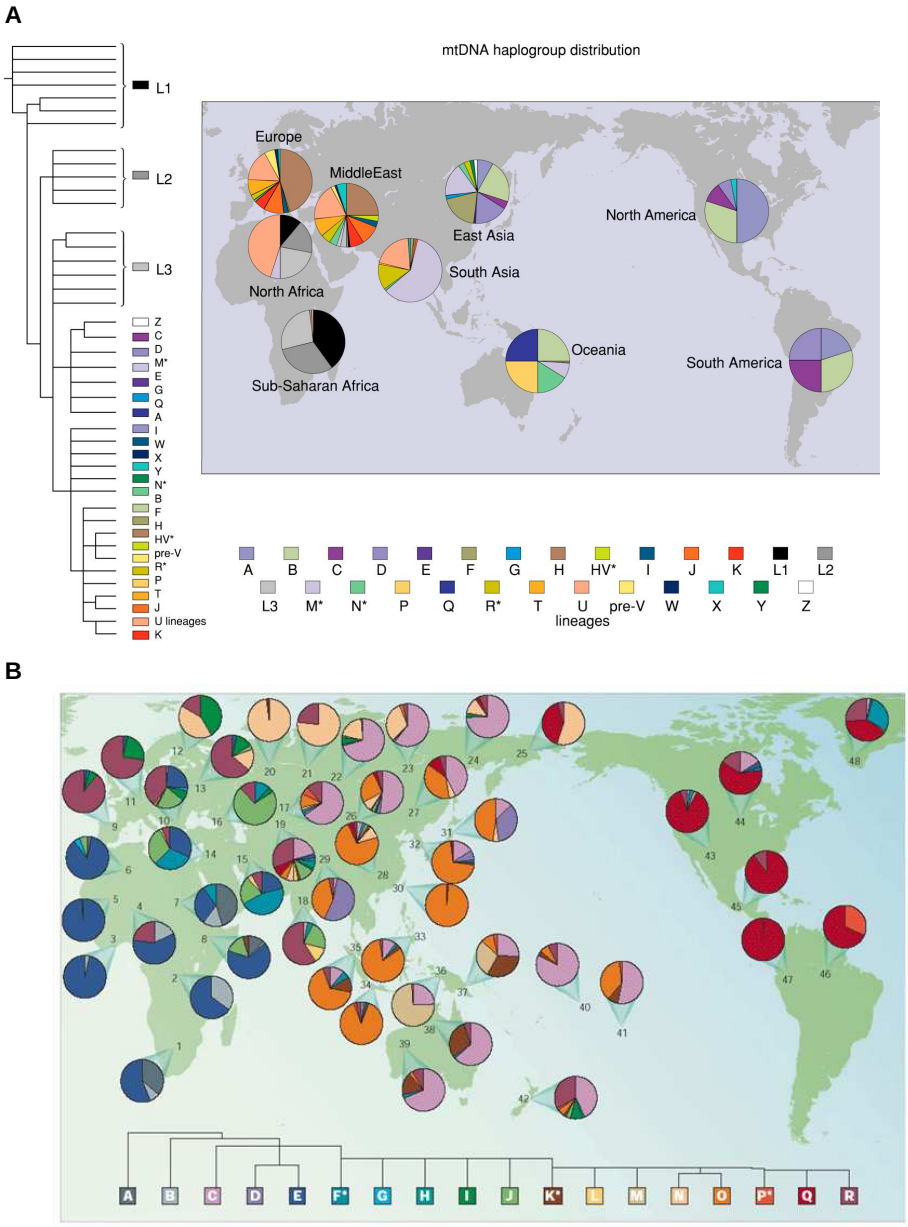


Figure 9. Geographical distribution of the major (A) mtDNA and (B) NRY clades. Each major clade (haplogroup) is assigned a colour reflecting its position in the phylogeny, and its frequency in population samples from broad geographical regions is shown in the pie charts. Respectively from Jobling et al. (2004) and Jobling and Tyler-Smith (2003).

1.3. Inferring adaptation from genomic variation

The investigation of natural selection and genetic adaptations in humans holds a central place in anthropology, human genetics, and evolutionary biology. In the past few years, we have witnessed an explosion of surveys of recent and ongoing natural selection. The development of large-scale genotyping technologies has generated an unprecedented amount of genetic data that allowed the scan of the entire human genome for signals of adaptation without the need of previous biological hypotheses. Hence, an important motivation for providing detailed selection maps stems from the idea that inferences about selection can provide important functional (phenotypic) information.

As humans colonized most of the world's landmasses, they encountered diverse habitats including savannas, forests, tundra, and deserts. The new environments imposed adaptive challenges regarding temperature, diet, and altitude among others. In addition, some cultural revolutions such as agriculture and cattle farming implied strong selective pressures for many human populations that recently (~10 kya) changed their ancestral lifestyle as hunter-gatherers. The study of targets of positive selection has generated much excitement in light of the possibility of finding the genetic basis of human uniqueness and gaining knowledge of population-specific adaptive phenotypes.

1.3.1. Methods for detecting classical selective sweeps

As a new advantageous mutation is driven to fixation because of positive selection, it leaves a footprint known as "selective sweep" or "hard sweep". This signal is characterized by a drastic reduction in variation and an increase in linkage disequilibrium as the beneficial allele drags out the linked neutral variants to high prevalence (Mainard Smith and Haigh 1974) (see Figure 10A). With time, and due to the action of recombination, the effect on the

strength of linkage disequilibrium will diminish at larger distances from the beneficial mutation, being stronger near the selected allele. Eventually, new mutations will restore the diversity lost by the "hitchhiking" event, producing a skew of the site frequency spectrum towards rare alleles.

Detecting the signature of natural selection essentially consists of distinguishing the patterns of variation shaped solely by demographic history from those that are also influenced by natural selection. The assumption is that most polymorphisms are neutral and governed by the stochastic effects of genetic drift (in agreement with the neutral theory), whereas positive selection is a locus-specific force. Because of its simplicity (it does not relay on neutral simulations based on models with uncertain reliability), the use of genome-wide empirical distributions has become very common in positive selection scans (Akey et al. 2002). Nevertheless, certain demographic scenarios (such as changes in population size and population subdivision) can produce confounding signals that difficult the identification of selective sweeps. Therefore, the power to detect selective sweeps depends directly on its strength (which determines the speed in frequency increase) and the time elapsed since the selective event.

Different genetic methods have been developed to detect selective sweeps at intra-species level (schematized in Figure 10B):

a. Long range haplotype methods. These methods search for recent and strong sweeps by identifying frequent haplotypes with high homozygosity extending over large regions. The most widely used statistics for incomplete sweeps are the extended haplotype test (EHH) (Sabeti et al. 2002) and the derivative integrated haplotype score (iHS) (Voight et al. 2006). To detect complete sweeps in inter-population comparisons the most used method is the Cross Population Extended Haplotype Homozygosity test (XP-EHH) (Sabeti et al. 2007).

b. Tests based on population differentiation. Positive selection can create allele frequency differences between subpopulations under different selective pressures. When a locus shows extraordinary levels of substructure compared to the genome-wide baseline, this can be interpreted as evidence for positive selection (Lewontin and Krakauer 1973). For instance, Akey and colleagues (2002) proposed candidate genes that showed extreme F_{ST} values with respect to the distribution of the statistic genome-wide.

c. Tests based on the frequency of derived alleles. If neutral, derived alleles are expected to be found at lower frequencies than the ancestral ones (Watterson and Guess 1977). In a selective sweep, derived alleles that are linked to the beneficial mutation can be hitchhiked to high frequency.

d. Site frequency spectrum (SFS) based methods. These methods aim to detect the loss of diversity in the vicinity of the beneficial allele (at high frequency) and an excess of new mutations that will accumulate with time. Several classical methods are based on the skew of the site frequency spectrum including Tajima's D (Tajima 1989) and Fu and Li's D* (Fu and Li 1993). Interestingly, the reduction of diversity lasts longer in time than other types of signatures (referred as "Heterozygosity/rare alleles" in Figure 10B). However, these tests are more vulnerable to the confounding effect of demography (i.e. an expanding population increases the fraction of rare alleles).

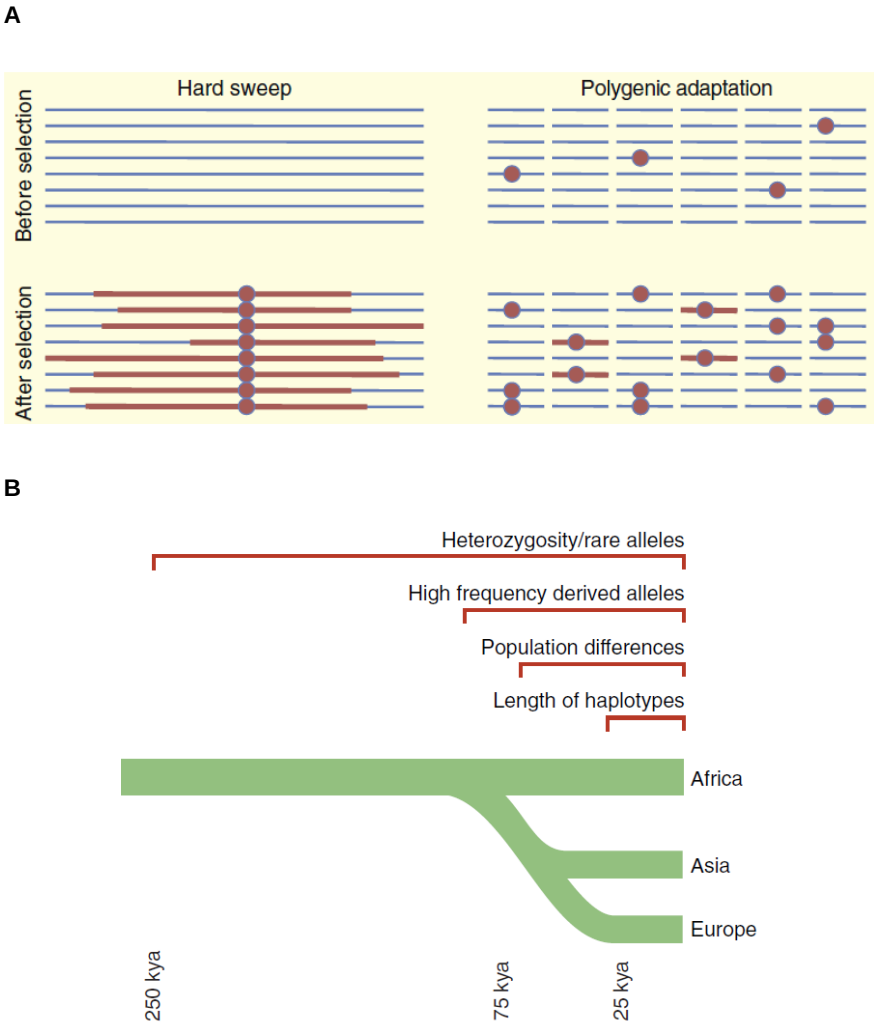


Figure 10. Signatures of selective sweeps. A. Hard and soft sweep. The horizontal blue lines represent haplotypes, the red lines indicate regions that are identical by descent (IBD) and the red circles indicate alleles that are favoured. In the monogenic model (hard selection, left panel), selection drives a new mutation to fixation, creating a large IBD region. In the polygenic model, and prior to selection red alleles exist at modest frequencies at various loci across the genome. After selection, the genome-wide abundance of favoured alleles has increased, but in this cartoon they have not fixed at any locus. In this example, at some loci selection has acted on new variants, creating signals of partial sweeps at those loci. From Pritchard et al. (2010). **B.** Time scales for the signatures of selection. A rough estimate of how long each signature is useful for detecting selection in humans is shown. Adapted from Sabeti et al. (2006).

1.3.2. Adaptation beyond classical sweeps

The methods described previously have allowed the identification of some interesting cases of adaptation in humans. For instance, selected variants in the lactase gene *LCT* confer the ability to digest milk in adults of herding European and African populations (Bersaglieri et al. 2004; Tishkoff et al. 2007), and the *EPAS* gene is involved in the recent adaptation to high altitude in Tibetans. However, it has been recently shown that few adaptations in humans actually occurred under the classical sweep model (Hernandez et al. 2011). This suggests that the well-known cases of positive selection are among the scant low-hanging fruits reachable by current methods based on hard sweeps. Therefore, most adaptations must have occurred under other modes of "soft sweeps", such as selection on standing variation or polygenic selection. Indeed, classical models in artificial and natural selection in the quantitative genetics literature are based on polygenic adaptation. In addition, genome-wide association studies (GWAS) have highlighted the extreme polygenic basis of many human traits. Adaptation through "soft sweeps" can be very efficient without large allele frequency shifts in each locus, without leaving the sweep footprint (Pritchard et al. 2010) (see Figure 10A). Therefore, the signature of adaptation in highly polygenic traits may be difficult to distinguish from stochastic signals. This will probably require the incorporation of phenotypic and/or environmental variables to the genetic analyses (Pritchard and Di Rienzo 2010).

1.4. Human genetic variation

Our species shows relatively less genetic variation than other primates, and the genetic differences found among human populations are also low (Fischer et al. 2006). These patterns of variation are explained by our species' origins and dispersal described below. Also in this section, the demographic histories of three human groups analyzed in the present thesis are introduced: Cubans, European Romani, and African Pygmies. Finally, the populations and data available in public catalogues of human genetic variation will be described.

1.4.1. Origin of Anatomically Modern Humans

The fossil record supports that the transition to anatomically modern *Homo sapiens* occurred in Africa and that by ~200,000 years ago the basic modern morphology was already established (Tattersall 2009). The most accepted model in the last decades is known as the "Recent African Replacement model" which posits that modern humans evolved in Africa and then spread and replaced all other existing hominins. The extreme alternative model, known as the "Multiregional model", proposes that the transition from *erectus* to *sapiens* took place independently in different places in the Old World in different times. Genetic data accumulated since the strong debate between the two models in the 80s has consistently supported the recent African model. The genetic evidence can be summarized in higher levels of diversity, weaker LD structure and more private alleles in Africans (Conrad et al. 2006; Jakobsson et al. 2008; Tishkoff et al. 2009), whereas non-Africans show a subset of the African diversity that is partitioned geographically reflecting serial founding events (Ramachandran et al. 2005).

Two recent landmark papers describing the genomes of extinct hominins, the Neanderthals (Green et al. 2010) and Denisovans (Reich et al. 2011), led to the exciting discovery that modern humans have admixed with other

species from the genus *Homo*. Specifically, the genomes of non-African human populations harbour around 1-4% of Neanderthal ancestry, whereas Near Oceanians (New Guineans and Bougainville Islanders) received an extra 4-6% of their genomes from Denisovans (Figure 11). In addition, further admixture events with unknown hominins have been suggested. For instance, some genomic features of African hunter-gatherers are better explained by archaic admixture with an African hominin around 35 kya (Hammer et al. 2011). Apart from challenging an strict African replacement model of the origin of our species, these findings also open the possibility that some adaptations in humans were due to genes acquired from our archaic relatives (Abi-Rached et al. 2011).

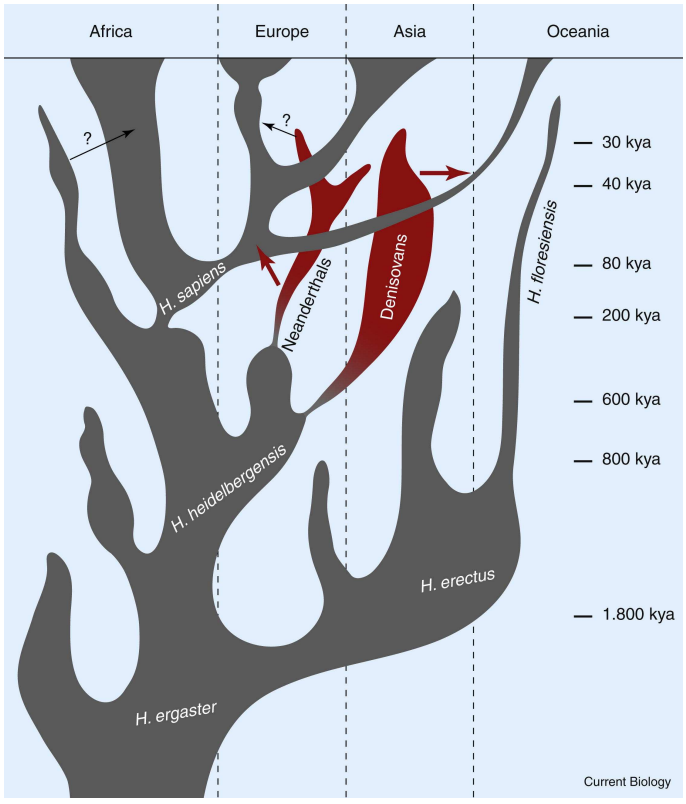


Figure 11. Hypothetical evolutionary relationships among modern humans, Neanderthals, and Denisovans. Red arrows mark genetic evidence of interbreeding among different hominin populations. Black arrows mark suggested or possible additional gene flow. From Lalueza-Fox and Gilbert (2011).

1.4.2. The peopling of Cuba

Human migrations during the colonial times dramatically changed the genetic landscape in some areas of the world. America is one of those regions, where Native Americans, European colonizers and African slaves have contributed to the current gene landscape of the continent.

Two different Native American groups were settled in Cuba when the Spaniards arrived in 1492: Ciboney hunter-gatherers and agriculturalist Tainos (Dacal-Moure and Rivero de la Calle 1986). The ancestors of both groups are believed to have arrived in the island around 7 and 3 kya respectively, in two migration waves originated in the Orinoco Valley in South America (Lalueza-Fox et al. 2003). As in other regions in the Caribbean, the arrival of Europeans triggered a fast and drastic demographic decline of the Native peoples in Cuba.

The slave trade to Cuba began by the 16th century and lasted until the Spanish Abolition Law in 1880. The estimates on the total number of African slaves brought to the island ranges between 700,000 (Curtin 1969) and 1,300,000 (Pérez de la Riva 1979). Historical records on the African regions where these peoples where enslaved are scarce but mainly point at West and South-east Africa. During these centuries, Spaniard immigration was also constant, especially from the Canary Islands.

Although the admixture process from populations originated in different continents is obvious in the extant Cuban population, the amount of admixture and the possible bias in contribution between males and females have not been assessed so far. In addition, although the Native Cuban populations went extinct soon after the arrival of the Europeans, the possibility of finding native genes in the current Cubans remains open.

1.4.3. The demographic history of European Romani (Roma)

The Romani are the largest minority group in Europe with more than 10 million people according to the Council of Europe (Roma and Travellers Division 2010). However, the censuses on the Romani people are usually considered to underestimate the real number. Because of their socio-economically marginalized condition, many countries are reluctant to recognize the Romani as an ethnic group. At the same time, the Romani usually avoid self-identification to avert stigmatization. There are different terms that are used for the Romani: 'Gyspy', 'Gitano', 'Cigano', 'Ijito', 'Sinti', 'Romanichal' etc. The terms 'Roma' or 'Romani' are preferred, as others can be considered derogatory or do not include all Romani groups.

The Romani society is built on 'groups' that present strong identity and rules of endogamy. These are based on several complex factors such as traditions, trades, languages and religions. As a consequence of the high degree of socio-cultural differentiation among Romani groups, the identification as Romani depends ultimately on self-assignation (Hancock 2002).

The Romani lack written historic or genealogic records that testify their origins. Therefore, historic, linguistic, and genetic sources are especially valuable to reconstruct the history of the Romani. Linguistic studies first located the origin of the Romani in the Indian continent (Fraser 1992; Liégeois 1994). According to these surveys, the proto-Romani speaker population left India between 5th and 10th centuries and moved westwards through the Indus Valley, Persia, Armenia, and Turkey before arriving at Europe (Fraser 1992). The first record of the Romani in Europe locates these itinerant peoples in the Thrace area (currently divided in Bulgaria, Greece and Turkey) in the 11th century (Fraser 1992). By the 14th century, the Romani were widely established in the Balkan region, rapidly spreading throughout all Europe during the next century (see Figure 12).

Genetic evidence confirmed the origin of the European Romani in India (Gresham et al. 2001). This study based on uniparental markers found that 45% of the paternal and 26% of the maternal haplogroups could be ascribed to the Indian subcontinent (Gresham et al. 2001). In addition, the low diversity and high internal genetic differentiation found suggested episodes of strong drift due to founding events and strong endogamy. Finally, the finding of non-Romani European lineages in the Romani indicated admixture with European hosts.

Figure 12. Proposed migration route of the Romani from India to Europe. Dots indicate the cities where the Romani were first recorded. Modified from Kenrick (2007).

As observed in other isolated founder populations, the Romani show high incidences of genetic diseases that are rare in surrounding populations. Several studies have shown that certain Mendelian diseases (some of them previously unknown, see Table 2) are caused by few mutations that affect many individuals from different geographic areas, strongly indicating a common origin and a founder effect (Kalaydjieva et al. 2001).

Disorder	OMIM*	Inheritance	Map Location	Gene	Mutation
Primary congenital Glaucoma	231300	AR	2p21	CYP1B1	E387K
Galactokinase Deficiency	230200	AR	17q24	GK1	P28T
Polycystic kidney Disease	173900	AD	4q21-q23	PKD2	R306X**
Hereditary motor and Sensory neuropathy-Lom	601455	AR	8q24	NDRG1	R148X
Hereditary motor and Sensory neuropathy-Russe	605285	AR	10q23		
Congenital cataracts facial dysmorphism neuropathy	604168	AR	18qter		
Limb girdle muscular dystrophy type 2C	253700	AR	13q12	SGCG	C283Y
Congenital myasthenia	254210	AR	17p13	CHRNE	1267delG
Glanzmann Thrombasthenia	273800	AR	17q21	ITGA2B	IVS15DS, G-A+1

Table 2. Mendelian disorders in the Romani caused by private founder mutations. From Kalaydjieva (2001).

Overall, genetic studies have been very useful to confirm the Indian origins of the Romani. However, details on the origin of the Romani within India remain unclear. Neither the geographic location of the parental Indian population nor the timing nor the magnitude of the founder event have been unraveled. In addition, the admixture relationships between each Romani population and the current (European) and past (Middle Eastern) host populations have not been fully explored. Finally, the fine-scale genomic description of the Romani has not been attempted, which could unravel the implications of the demographic history and social customs in their disease status.

1.4.4. Demographic and adaptive history of African Pygmies

The term "Pygmy" is used in anthropology to refer to several worldwide populations that show average adult male height below 155 cm (Cavalli-Sforza 1986). Despite the word 'Pygmy' could be considered pejorative, there is no alternative term to include the, at least, 15 different ethno-linguistic groups (around 300,000 to 500,000 people, see Figure 13) that live on hunter-gathering in the Central African rainforest (Ohenjo et al. 2006).

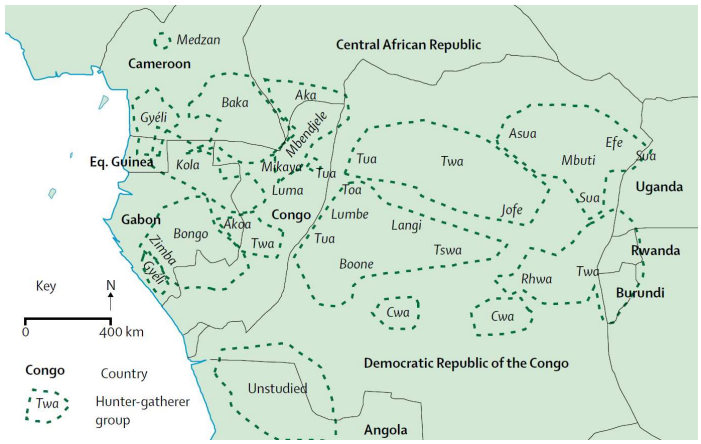


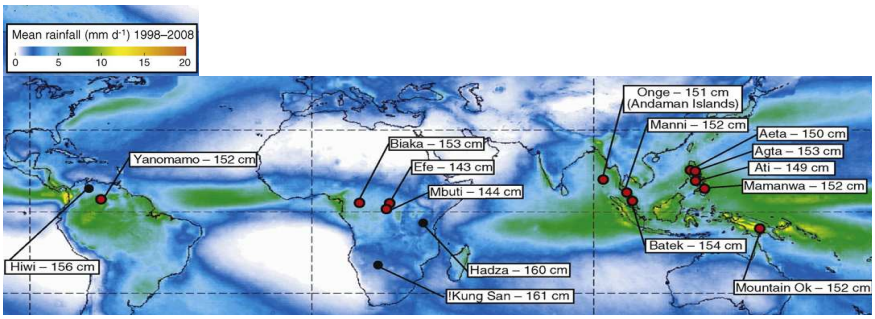
Figure 13. Pygmy peoples of Central Africa. From Ohenjo (2006).

African Pygmies share the territory with non-Pygmy agriculturalist groups, mostly Bantu speakers. The Pygmy and their neighbour agriculturalist groups maintain symbiotic relationships based on the exchange of forest products for cultivated foods and iron tools (Sayer et al. 1992). The contact between the two groups is believed to have existed for the last 4 ky (Cavalli-Sforza 1986). During this time the Pygmies adquired their neighbour's language (Bantu, Sudanic or Adamawa-Ubangian). Even though intermarriages between Pygmy and non-Pygmies exist, these are mostly limited to Pygmy females and non-Pymgy males (Cavalli-Sforza 1986; Quintana-Murci et al. 2008).

Genetics has been revelatory in disentangling the population history of Central Africa, where archaeological evidence are scarce. Together with the Southern and Eastern African Khoisan, Central African Pygmies are located at the root of the tree of all human populations (Jakobsson et al. 2008; Li et al. 2008). Recent genetic studies based on autosomal and mtDNA data have revealed that the common ancestor of the Pygmies and Bantu agriculturalist split around 60 kya and that the Western and Eastern Pygmy groups separated around 20 kya (Batini et al. 2011; Patin et al. 2009).

Why are Pygmies small? Anthropologists have recurrently addressed this question in the last 20 years. Yet, the underlying genetic and physiological factors, as well as the evolutionary forces that act on this trait remain unknown. The Pygmy phenotype is associated with hunter-gatherer populations that inhabit the tropical rainforests not only in Africa, but also in America and Asia (see Figure 14A). This association has been interpreted as adaptation to tropical rainforests. It has been argued that a decreased size would involve less heat production and a more effective thermoregulation (Cavalli-Sforza 1986). Other explanations relate to a more efficient mobility in dense forests (Diamond 1991) and less caloric demand (Shea and Bailey 1996). More recently, a life-history based study suggested that the Pygmy stature could be consequence of strong selective pressures acting on an earlier onset of reproduction (Migliano et al. 2007). The high mortality rates in the Pygmies would overcome the detrimental effects of early growth cessation and the consequent short stature. In any case, the convergent evolution of the Pygmy phenotype in different continents with similar ecosystems suggests that the phenotype is strongly adaptive (Perry and Dominy 2009). Possibly, the identification of the genetic variants that participate in the genetic architecture of the phenotype will help elucidating its putative adaptive value.

A



B

Figure 14. African Pygmy phenotype. A. Association of the human Pygmy phenotype with tropical rainforest habitats. Approximate locations of small-bodied hunter-gatherer populations, with mean adult male stature estimates. The smallest modern human statures (mean adult male height < 155 cm) are always associated with tropical rainforests (red circles). Some hunter-gatherer populations occupying savanna-woodlands (black circles) are also relatively small, such as the Hiwi of the Venezuelan llanos, the Hadza of Tanzania and the !Kung San of Botswana and Namibia. Precipitation data are from the Tropical Rainfall Measuring Mission (Goddard Space Flight Center, National Aeronautics and Space Administration; http://trmm.gsfc.nasa.gov). **B.** Efe male, Democratic Republic of Congo. Modified from Perry and Dominy (2009).

1.4.5. Public resources of human DNA variation

The considerable amount of genomic data accumulated in the last years provides an excellent source to study current and past demographic patterns as well as adaptive histories in several human populations. In this section I will introduce three public resources: the HGDP, HapMap, and 1,000 Genomes Project.

The objective of the Human Genome Diversity Project (HGDP, (Cavalli-Sforza et al. 1991)) was to provide with a good representation of native populations worldwide (many of them isolates) to study human evolution. An important achievement of the HGDP Project was the establishment of the Human Genome Diversity Cell Line Panel, HGDP-CEPH (Cann et al. 2002), consisting of cultured lymphoblastoid cell lines from 1,050 individuals in 52 populations from the five continents (see populations in Figure 15). The HGDP collection was the first worldwide human DNA catalogue that was available to not-for-profit researchers. These samples have been included in numerous genetic studies on dense STR (Rosenberg et al. 2005; Rosenberg et al. 2002) and SNP data (Jakobsson et al. 2008; Li et al. 2008) which are publicly accessible and have been used in this thesis.

The primary objective of the International HapMap Project (The International HapMap Consortium 2003) was to develop a haplotype map of the human genome to study common patterns of human DNA sequence variation. The description of the haplotype structure was aimed to provide the basis of LD based association studies for complex diseases (Goldstein and Weale 2001). The publicly available HapMap dataset has become a very useful source to study human demography and selection (Barreiro et al. 2008). After the second phase, over 3.1 million SNPs were characterized in 270 individuals from four populations (European Americans, Japanese, Han Chinese and Yoruba from Nigeria). In addition, in the last release (HapMap

3) 1.6 million SNPs were genotyped in a total of 1,184 individuals from 11 global populations, including admixed populations such as the Mexican "mestizos" and African Americans (see Figure 15).

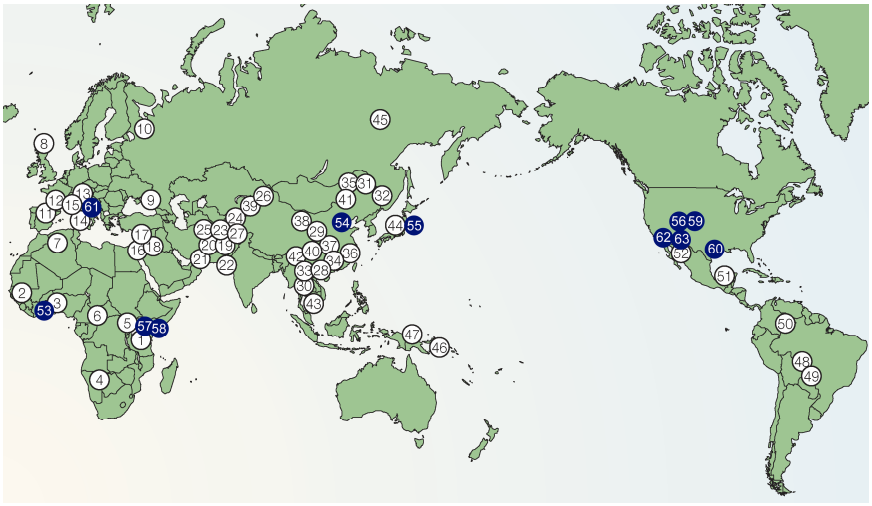


Figure 15. HGDP-CEPH and HapMap sample locations. White and blue circles respectively. 1-Bantu, 2-Mandenka, 3-Yoruba, 4-San, 5-Mbuti, 6-Biaka, 7-Mozabite, 8-Orcadian, 9-Adygei, 10-Russian, 11-Basque, 12-French, 13-North Italian, 14-Sardinian, 15-Tuscan, 16-Bedouin, 17-Druze, 18-Palestinian, 19-Balochi, 20-Brahui, 21-Makrani, 22-Sindhi, 23-Pathan, 24-Burusho, 25-Hazara, 26-Uygur, 27-Kalash, 28-Han (S. China), 29-Han (N.), 30-Dai, 31-Daur, 32-Hezhen, 33-Lahu, 34-Miao, 35-Oroqen, 36-She, 37-Tujia, 38-Tu, 39-Xibo, 40-Yi, 41-Mongola, 42-Naxi, 43-Cambodian, 44-Japanese, 45-Yakut, 46-Melanesian, 47-Papuan, 48-Karitiana, 49-Surui, 50-Colombian, 51-Maya, 52-Pima, 53-Yoruba (YRI), 54-Han (CHB), 55-Japanese (JPT), 56-USA European (CEU), 57-Luhya (LWK), 58-Maasai (MKK), 59-USA Chinese (CHD), 60-USA Gujarati (GIH), 61-Toscani (TSI), 62-USA Mexicans (MXL), 63-USA Africans (ASW). Modified from Cavalli-Sforza (2005).

The design of the HapMap SNP discovery had important consequences on different aspects of genetic data. The SNPs were first identified by direct sequencing of a small panel of individuals and subsequently genotyped in a larger panel. Therefore, and given that the probability of discovery of an SNP depended on its allele frequency, this procedure prioritized common alleles introducing a substantial bias in the site frequency spectrum. This ascertainment bias results in the overestimation of heterozygosity and population subdivision in HapMap data (Clark et al. 2005). Importantly, this

needs to be considered when trying to do population genetic inferences with HapMap data or any commercial SNP arrays based on the HapMap SNPs, with special emphasis in certain populations (Figure 16).

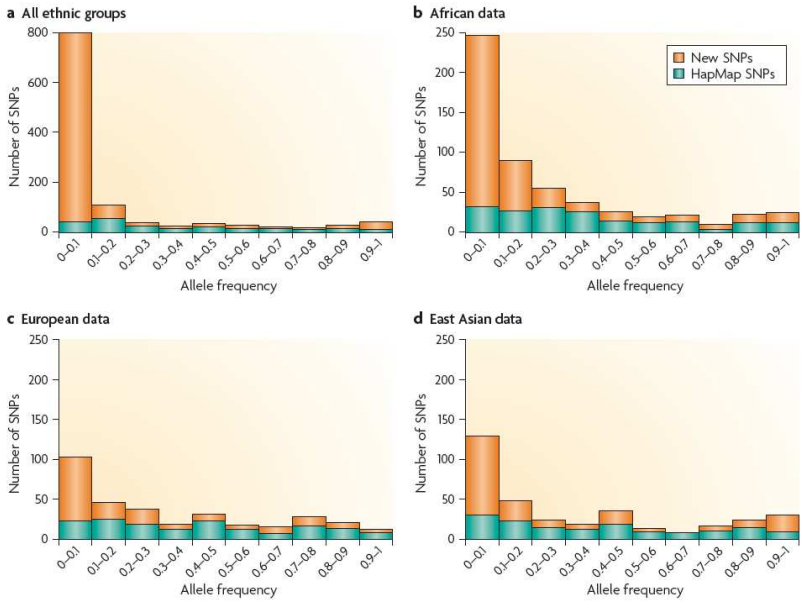


Figure 16. Ascertainment bias in SNP databases. Number of SNPs in the HapMap data (green) compared with those discovered by resequencing 40 intergenic regions in 90 individuals from 6 ethnic groups (Wall et al. 2008) and that were not present in the HapMap data (orange), categorized by derived allele frequency. **a.** Data from all ethnic groups combined. **b.** SNPs discovered in Mandinka compared with Yoruba (YRI).**c.** SNPs discovered in Basque compared with CEU. **d.** SNPs discovered in Han Chinese compared with Han (CHB). From Teo et al. (2010).

The "successor" of the HapMap Project is the 1,000 Genome Project, launched at 2008 and expected to be finished by 2012. This project aims to fully sequence the genomes of 2,500 individuals from 27 populations, so that the full spectrum of human polymorphism can be described. The project has been encouraged by the possibility that rare variants (mostly population specific) could be important to explain the "missing heritability" of complex diseases (Bustamante et al. 2011). Undoubtedly, complete resequencing data will also constitute an important resource to study human evolution, free of ascertainment bias (Gravel et al. 2011).

2. Objectives

The objective of this work is to use genetic diversity data to learn about past demographic events and the history of natural selection in some human populations. Specifically:

The information provided by uniparental markers in the Cuban population is used to:

1. Identify the geographic origin of the ancestors of current Cubans
2. Determine if genes of Native Cubans persist in Cuban genetic legacy
3. Quantify the genetic contribution of each parental population
4. Study sexual asymmetry in the contribution of parental populations

Genome-wide data in the European Romani is surveyed with the aim to:

5. Locate and date the origin of the European Romani in the Indian subcontinent
6. Quantify the magnitude of the founder event
7. Study the migration route followed by the European Romani
8. Quantify the admixture between European Romani and European hosts
9. Determine fine-scale population structure of the European Romani
10. Study the impact of the demographic history and social customs of the Romani on the prevalence of Mendelian and complex diseases

The genome-wide patterns of African Pygmy and non-Pygmy populations are analyzed with the objective to:

11. Gain insights on the genetic architecture of the Pygmy height by developing a new statistic that incorporates phenotypic information
12. Provide with a list of candidate genomic regions with strong evidence for having a role in the trait
13. Model the demographic history of the African Pygmies to assess the role of natural selection in the evolution of the candidate regions

3. Results

3.1. Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba

Isabel Mendizabal, Karla Sandoval, Gemma Berniell-Lee, Francesc Calafell, Antonio Salas, Antonio Martínez-Fuentes, and David Comas

BMC Evol Biol. 2008 Jul 21;8(1):213

3.2. Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective

Isabel Mendizabal, Cristina Valente, Alfredo Gusmaõ, Cíntia Alves, Verónica Gomes, Ana Goios, Walther Parson, Francesc Calafell, Luis Alvarez, António Amorim, Leonor Gusmaõ, David Comas, and Maria João Prata

PLoS One. 2011 Jan 10; 6(1):e15988

3.3. Reconstructing the population history of European Romani from genome-wide data

Isabel Mendizabal, Oscar Lao, Urko M. Marigorta, Andreas Wollstein, Leonor Gusmão, Vladimir Ferak, Mihai Ioana, Albena Jordanova, Radka Kaneva, Anastasia Kouvatsi, Vaidutis Kučinskas, Halyna Makukh, Andres Mestpalu, Mihai G. Netea, Rosario de Pablo, Horolma Pamjav, Dragica Radojkovic, Sarah J.H. Rolleston, Jadranka Sertic, Milan Macek Jr., David Comas, and Manfred Kayser

Submitted

Reconstructing the population history of European Romani from genome-wide data

Isabel Mendizabal[1,*], Oscar Lao[2,*], Urko M. Marigorta[1], Andreas Wollstein[2], Leonor Gusmão[3,4], Vladimir Ferak[5], Mihai Ioana[6,7], Albena Jordanova[8,9], Radka Kaneva[9], Anastasia Kouvatsi[10], Vaidutis Kučinskas[11], Halyna Makukh[12], Andres Mestpalu[13], Mihai G. Netea[14,15], Rosario de Pablo[16], Horolma Pamjav[17], Dragica Radojkovic[18], Sarah J.H. Rolleston[19], Jadranka Sertic[20,21], Milan Macek Jr.[22], David Comas[1,#,$], and Manfred Kayser[2,#,$]

[1]Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Spain

[2]Department of Forensic Molecular Biology, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands

[3]IPATIMUP – Institute of Pathology and Molecular Immunology of the University of Porto, Porto, Portugal

[4]Medical and Human Genetics Laboratory, and Molecular Biology and Genetics Post-graduate Program, Federal University of Pará (UFPA), Belém, Pará, Brazil

[5]Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia

[6]University of Medicine and Pharmacy Craiova, Craiova, Romania

[7]University of Medicine and Pharmacy Carol Davila Bucharest, Bucharest, Romania

[8]VIB Department of Molecular Genetics, University of Antwerp, Antwerp, Belgium

[9]Department of Chemistry and Biochemistry, Molecular Medicine Center, Medical University Sofia, Sofia, Bulgaria

[10]Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece

[11]Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Vilnius, Lithuania

[12]Institute of Hereditary Pathology of the Ukrainian Academy of Medical Sciences, Lviv, Ukraine

[13]Estonian Genome Center, University of Tartu, Tartu, Estonia

[14]Department of Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

[15]Nijmegen Institute for Infection, Inflammation and Immunity, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

[16]Servicio de Inmunología, Hospital Universitario Puerta de Hierro, Madrid, Spain

[17]DNA Laboratory, Institute of Forensic Medicine, Network of Forensic Science Institutes, Budapest, Hungary

[18]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia

[19]Institute of Medical Genetics, University Hospital of Wales, Cardiff, Wales, United Kingdom

[20]Clinical Institute of Laboratory Diagnosis, Zagreb University Hospital Centre, Zagreb Croatia

[21]Department of Chemistry, Biochemistry and Clinical Biochemistry, School of Medicine, University of Zagreb, Zagreb, Croatia

[22]Department of Biology and Medical Genetics, University Hospital Motol and the 2nd Faculty of Medicine, Charles University, Prague, Czech Republic

* These authors contributed equally to this work

# These authors contributed equally to this work

$ Correspondence: david.comas@upf.edu or m.kayser@erasmusmc.nl

Summary

The Romani, the largest minority group in Europe with approximately 11 million people [1], constitute a mosaic of languages, religions and lifestyles while strong social organization makes them identifiable from other European peoples. Early linguistic [2] and recent genetic studies [3-7] have located the Romani origins in India. However, a genome-wide perspective on Romani origins, their fine-scale population structure and a detailed reconstruction of their demographic history are yet to be provided. Here we show by using genome-wide data from 13 Romani groups collected across Europe that the Romani Diaspora constitutes a single founder population which originated in north-western India ~1.1 thousand years ago (kya). We demonstrate that the initial founder event from India and secondary bottlenecks during their European dispersal from the Balkans (estimated ~0.7 kya) together with concurrent genetic isolation and differential admixture with non-Romani Europeans, all account for the strong population substructure and high levels of homozygosity observed within the European Romani. Inferred patterns of gene flow with non-Romani Europeans indicate that the Romani endogamous practices varied through space and time. Furthermore, we show that the Indian origins of the European Romani may play a role in their differential prevalence of complex diseases, and that their unique demographic history increases the risk of potentially harmful alleles. Overall, our findings help to complete the demographic history of Europe and highlight medical implications of the Romani genetic heritage.

Results and Discussion

Recently, the fine-scale genetic substructure of Europeans has been deciphered by use of genome-wide data [8-10]; however, these studies did not include the Romani, the largest minority group in Europe. Furthermore, important details of the Romani population history such as the location, dating and magnitude of the Out-of-India event, or their relationships with other populations remain elusive thus far. Such an exhaustive genetic reconstruction is not only relevant for a better understanding of the demographic history of the Romani people, and of Europeans in general, but also to evaluate the potential medical implications of the Romani demographic history. To address these issues we studied the genome-wide diversity of European Romani by analyzing ~800,000 single nucleotide polymorphisms (SNPs) in 152 individuals of 13 Romani groups collected from eastern, western and northern parts of Europe (supplementary Figure 1). We survey this data in the context of 3,313 individuals from non-Romani European [8], Indian [11], Middle Eastern [12, 13], and other worldwide populations [13, 14] (supplementary Table 1).

A first multidimensional scaling (MDS) plot based on identity-by-state (IBS) distances between pairs of worldwide individuals locates the European Romani within the west Eurasian cline (supplementary Figure 2). To disentangle the genetic influence of the populations located geographically between India and Europe, we estimate the proportion of ancestral populations for each individual using ADMIXTURE [15]. At $k = 2$, a genetic continuum from India to Europe is observed, where the Indian ancestral component dilutes westwards but appears higher in the Romani than in non-Romani Europeans and the Middle Easterners (supplementary Figure 3). An additional ancestral component is observed at $k = 3$ in Pakistani, Iraqi Kurds, Qatari and individuals from the Caucasus as well as in southern non-Romani Europeans, but not in the European Romani. This finding may indicate either that the sampling coverage of the actual Romani migration route is incomplete in our study, or that the European Romani received considerable linguistic but not genetic influx from populations in the Near and Middle East. The latter is in agreement with the lack of historical records testifying a long presence of the Romani in the Middle East [2], and would

suggest a rapid rather than gradual migration of the Romani people from India to the region of the Byzantine Empire.

After discarding a major genetic influence from the Middle East with the previous analyses, we built a second MDS considering only the European Romani and the two putative parental populations i.e., non-Romani Europeans and Indians. The first dimension distinguishes non-Romani Europeans from Indian individuals, whereas the second dimension separates the European Romani from both of their putative parental groups (Figure 1). Although some European Romani individuals are dispersed towards non-Romani Europeans, indicative of recent genetic admixture; most European Romani cluster towards Indians, suggesting a single founding event with considerable genetic isolation. A Neighbor-Joining [16] (NJ) tree using Weir and Cockerham's F_{ST} distances [17] between pairs of populations confirms this finding. All European Romani groups cluster together and separately from Indians and non-Romani Europeans, except for the Welsh Romani who cluster with non-Romani Europeans from the British Isles (Figure 2). An Analysis of Molecular Variance (AMOVA [18]) shows that population branches are significantly longer within the European Romani than within the non-Romani Europeans and the Indians, reflecting strong genetic isolation among the European Romani (see legend Figure 2). The proportion of ancestral populations for each individual by ADMIXTURE at $k = 2$ identifies one ancestral component exclusively present in Indians and European Romani; the other component being shared between European Romani, non-Romani Europeans, and Indians (Figure 3). The presence of Western Eurasian ancestry in Indo-European and, to a lesser extent, Dravidian speaking Indian populations is in agreement with their demographic history [11]. At $k = 3$ (best k determined by means of cross-validation, supplementary Figure 5) a Romani-specific ancestral component encloses completely the Indian and partially the Western Eurasian components present in the European Romani. The median membership of the Romani component is above 60% (minimum 61% in Spain and maximum 99% in Ukraine), except again for the Welsh Romani, who show minimal Romani membership (ranging from 0 to 15%). In contrast, almost a quarter of all European Romani individuals show considerable non-Romani European component (above 30%), although the individual variance of this component is

high, suggesting recent genetic admixture. At higher *k*-s new Romani ancestral components arise, which are more frequent in the Balkan Romani (supplementary Figure 6 for more *k*-s). Notably, the diversity in these Romani ancestral components within each population decays linearly with geographic distance from the Balkans (supplementary Figure 7). Following a serial founder effect colonization model [19], this pattern supports the Balkan area as the source of all European Romani. Moreover, these new ancestral components reveal strong genetic affinities among geographically dispersed European Romani groups such as those from Central Europe, the Baltic region, and from Iberia.

To further confirm these affinities, we ran HAPMIX [20] to identify the segments of Indian and non-Romani European origin in the European Romani genomes. The analyses based on SNPs of Indian ancestry (see MDS and NJ tree in supplementary Figure 8), thus avoiding the genetic influence of non-Romani Europeans, confirm that the western (Spanish and Portuguese) and northern (Estonian, Lithuanian) European Romani shared the Out-of-Balkans migration route. Therefore, and in contrast to the strong correlation between geography and genetic diversity observed in non-Romani Europeans [8, 9], the migration route is the major determinant of the genetic affinities among the European Romani. Additionally, we use the non-Romani European segments identified by HAPMIX to study the temporal dynamics of admixture between the European Romani and non-Romani Europeans. Recent admixture is expected to produce individuals with long tracks of the non-Romani European component as well as others with no traces of admixture in the same population, while with time recombination will shorten and spread these admixed chromosomes across the population. We detect low presence of non-Romani European ancestry in the Balkan Romani but also long fragments in some Romani individuals from Bulgaria and Croatia (supplementary Figure 9), indicating a recent and ongoing shift in social rules. Oppositely, northern and western European Romani groups (such as those from Lithuania, Portugal, and Spain) show higher non-Romani European admixture but in shorter chromosomal tracks, suggesting older patterns of admixture in those European Romani that shared the out-of-Balkan migration.

We next performed analyses based on approximate Bayesian computation (ABC [21]) to determine the location and timing of the major events

in the Romani demographic history (see models tested in supplementary Figures 10 and 11). Our genome-wide data link the origin of the proto-Romani population with the Indo-European speaking populations from north-western India (Kashmiri Pandit, Tharu, and Meghawal account for 40% of the posterior probability). Notably, this highlighted geographic region in India encloses the Punjab area that was suggested previously as source of the Romani by anthropological, linguistic [22], and mtDNA [7] evidence (see supplementary Note 1). We further date the Out-of-India founder event circa 1.1 kya and identify a severe reduction of ~50% of the proto-Romani effective population size. Furthermore, we detect secondary bottleneck events during the dispersal of the Romani across Europe. Specifically, the Balkan and western European Romani (using Bulgaria and Spain as respective proxies) show a recent split around 0.7 kya, reducing the effective population size of these groups to only one third of the size of the parental Indian population (see Figure 4). These results, which are in agreement with historical records [2], point at cumulative drift during the serial founding events as one of the forces driving the extensive genetic differentiation observed in the European Romani, regardless of their recent origin.

Accordingly, the comparison of autozygosity levels among populations confirms the major Out-of-India and Out-of-Balkans founding events in the Romani history. In particular, we find more and longer homozygous chromosomal segments (runs of homozygosity, ROH) in the European Romani genomes compared to the ancestral Indian and non-Romani European genomes. The ROH patterns are also informative regarding social practices. Particularly, the pervasive presence of very large ROH tracks (>20Mb) indicates that consanguineous marriages are common in all European Romani groups (see supplementary Table 2 and supplementary Figure 10). A particular case is reflected by the Welsh Romani, who despite showing a predominant genome-wide non-Romani European ancestry also display extensive ROHs. The finding of typically Indian mtDNA lineages in Welsh Romani samples (data not shown) confirms their at least partial Romani origin. Thus, our data suggests that either the Welsh Romani admixed *in situ* with non-Romani Europeans and afterwards underwent strong isolation, or they received genetic admixture with an already isolated local population, such as the so called Native Travellers [23].

Since extensive homozygosity can result in a higher frequency of harmful recessive mutations, increasing the risk for both Mendelian and complex diseases [24], we aimed to evaluate possible medical implications due to the unique demographic history of the Romani. We first test whether there is an increase of homozygote genotypes for alleles in non-synonymous (ns) SNPs in the European Romani genome compared to their putative parental populations non-Romani Europeans and Indians. In contrast to the moderate enrichment (~1.1) found when the whole ns SNP spectrum is considered, we observe a three-fold excess of homozygotes in the European Romani for the rarest allele in the parental populations (MAF <0.05, see supplementary Note 2). To which extent these alleles represent a putative health threat is unknown. However, this result is consistent with purifying selection being less effective in removing slightly deleterious alleles during bottlenecks and founder events. Interestingly, several mainly recessive monogenic diseases caused by private mutations have been described in the Romani people [4]. Next, we analyze the frequency of SNPs associated with complex diseases by recent genome-wide association studies (supplementary Note 3). In four out of five studied cardiovascular/metabolic diseases, the average risk allele frequencies (RAF) of 202 associated SNPs is significant larger in the European Romani than in the non-Romani Europeans. In contrast, the same pattern of larger RAF as observed in the European Romani is also found in the Indians. Noteworthy, some genes involved in these traits show signs of recent positive selection in Indians [25]. It is tempting to speculate here that these disparities in metabolic disease prevalence may be explained by 'thrifty' genetic variants [26] of Indian origin that are detrimental under a western lifestyle. However, this result is preliminary as the bulk of heritability for these diseases remains undiscovered.

The present study constitutes the most comprehensive survey so far on the genetic characterization of the European Romani. Based on our findings, the genetic origins of the European Romani can be ascribed to the north-western Indian pool around 1.1 kya with a more recent contribution from non-Romani Europeans that varies both temporally and geographically. The founder and bottleneck effects together with endogamy explain the large levels of genetic differentiation and account for the excess of potentially harmful alleles in the

Romani. Intriguingly, the Indian origins of the Romani may be informative to understand the differences in complex disease prevalence compared to other Europeans. As this first genomic survey foresees, the current burst of medical genomics holds promising for improving the health status of the Romani people.

Experimental Procedures

Blood and buccal samples were collected with informed consent from unrelated volunteers who self-identified as Romani. All DNA samples were genotyped on Affymetrix 6.0 arrays. We performed the analyses with 152 samples that showed no signs of genetic relatedness and to 807,002 autosomal SNPs with no evidence for problematic genotyping. For some analyses we merged our data with Affymetrix 500K data from non-Romani European samples [8], Affymetrix 6.0 data from Indians [11] and Middle Eastern populations genotyped on Affymetrix 500K [12] and Affymetrix 250K [13], and worldwide populations on Affymetrix 250K [13] as well as with HapMap [14] data. This merged dataset contained 196,927 autosomal SNPs. Identity-by-state (IBS) distances and runs of homozigosity (ROH-s) were computed by PLINK v1.07 [27]. Multidimensional scaling (MDS) plots and correlation tests were computed by R [28]. Neighbor Joining trees [16] based on Weir and Cockerham´s F_{ST} distances [17] were drawn with the PHYLIP package 3.69 [29]. Analysis of Molecular Variance (AMOVA) [18] was performed with Arlequin version 3.11 [30]. Genetic ancestry proportions for each individual were estimated by ADMIXTURE software [15]. HAPMIX [20] was used to infer the chromosomal segments of non-Romani European and Indian ancestry in the Romani genomes. Competing demographic models were studied by Approximate Bayesian Computation [21] to infer demographic parameters. For more experimental and analysis details see supplementary materials.

Supplemental Information

Supplemental Information includes additional experimental procedures, results, and discussion.

Acknowledgements

We thank Jordi Camí and Francesc Valentí for their valuable help in collecting Romani samples from Spain, Natasa Petrovic for collecting Romani samples from Serbia, and Lazarus P. Lazarou for collecting Romani samples from Wales, United Kingdom. IM was supported by a PhD grant by the Basque Government (Hezkuntza, Unibertsitate eta Ikerketa Saila, Eusko Jaurlaritza, BFI107.4). OL, AW, and MK were supported by the Erasmus MC University Medical Center Rotterdam. UMM was supported by a PhD grant by Universitat Pompeu Fabra. MGN was supported by a Vici grant of the Netherlands Organization of Scientific Research. LG was supported by an Invited Professor grant from CAPES/Brazil. AM was supported by the Estonian Government grant SF0180142s08. This study was supported in parts by the Spanish Government MCINN grant CGL2010-14944/BOS to DC, the Czech Republic Ministry of Health grants CZ.2.16/3.1.00/24022 and 00064203 to MM, the Republic of Serbia Ministry of Education and Science grant ON173008 to DR, by the Belgium University of Antwerp grant IWS BOFUA 2008/23064 to AJ, and by the Portuguese Foundation for Science and Technology (FCT) project grant PTDC/ANT/70413/2006 to LG. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Education and Science and is partially supported by FCT.

References

1. Liegeois, J.-P. (1998). Roma, Gypsies, Travellers., (Strasbourg: Council of Europe Press.: Stationary Office Books).
2. Fraser, A. ed. (1992). The Gypsies (Oxford: Blackwell Publishers).
3. Gresham, D., Morar, B., Underhill, P.A., Passarino, G., Lin, A.A., Wise, C., Angelicheva, D., Calafell, F., Oefner, P.J., Shen, P., et al. (2001). Origins and divergence of the Roma (gypsies). American journal of human genetics *69*, 1314-1331.
4. Kalaydjieva, L., Gresham, D., and Calafell, F. (2001). Genetic studies of the Roma (Gypsies): a review. BMC Med Genet *2*, 5.

5. Gusmao, A., Gusmao, L., Gomes, V., Alves, C., Calafell, F., Amorim, A., and Prata, M.J. (2008). A perspective on the history of the Iberian gypsies provided by phylogeographic analysis of Y-chromosome lineages. Ann Hum Genet *72*, 215-227.

6. Kalaydjieva, L., Calafell, F., Jobling, M.A., Angelicheva, D., de Knijff, P., Rosser, Z.H., Hurles, M.E., Underhill, P., Tournev, I., Marushiakova, E., et al. (2001). Patterns of inter- and intra-group genetic diversity in the Vlax Roma as revealed by Y chromosome and mitochondrial DNA lineages. Eur J Hum Genet *9*, 97-104.

7. Mendizabal, I., Valente, C., Gusmao, A., Alves, C., Gomes, V., Goios, A., Parson, W., Calafell, F., Alvarez, L., Amorim, A., et al. (2011). Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective. PLoS ONE *6*, e15988.

8. Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., Comas, D., et al. (2008). Correlation between genetic and geographic structure in Europe. Curr Biol *18*, 1241-1248.

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

10. Nelis, M., Esko, T., Magi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskackova, T., Balascak, I., Peltonen, L., et al. (2009). Genetic structure of Europeans: a view from the North-East. PLoS ONE *4*, e5472.

11. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian population history. Nature *461*, 489-494.

12. Hunter-Zinck, H., Musharoff, S., Salit, J., Al-Ali, K.A., Chouchane, L., Gohar, A., Matthews, R., Butler, M.W., Fuller, J., Hackett, N.R., et al. (2010). Population genetic structure of the people of Qatar. American journal of human genetics *87*, 17-25.

13. Xing, J., Watkins, W.S., Witherspoon, D.J., Zhang, Y., Guthery, S.L., Thara, R., Mowry, B.J., Bulayeva, K., Weiss, R.B., and Jorde, L.B.

(2009). Fine-scaled human genetic structure revealed by SNP microarrays. Genome research *19*, 815-825.

14. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52-58.
15. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome research *19*, 1655-1664.
16. Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution *4*, 406-425.
17. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 13.
18. Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics *131*, 479-491.
19. Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., and Cavalli-Sforza, L.L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America *102*, 15942-15947.
20. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS genetics *5*, e1000519.
21. Tavare, S., Balding, D.J., Griffiths, R.C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. Genetics *145*, 505-518.
22. Hancock, I. (2002). We are the Romani people, (Hertfordshire: University of Hertfordshire Press).

23. Matras, Y. ed. (2010). fra (Edinburgh: Edinburgh University Press).

24. McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., et al. (2008). Runs of homozygosity in European populations. American journal of human genetics *83*, 359-372.

25. Metspalu, M., Romero, I.G., Yunusbayev, B., Chaubey, G., Mallick, C.B., Hudjashov, G., Nelis, M., Magi, R., Metspalu, E., Remm, M., et al. (2012). Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. American journal of human genetics *89*, 731-744.

26. Shah, A.M., Tamang, R., Moorjani, P., Rani, D.S., Govindaraj, P., Kulkarni, G., Bhattacharya, T., Mustak, M.S., Bhaskar, L.V., Reddy, A.G., et al. (2011). Indian siddis: african descendants with Indian admixture. American journal of human genetics *89*, 154-161.

27. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics *81*, 559-575.

28. R Development Core Team (1991). R: a language and environment for statistical computing. (Vienna: R Foundation for Statistical Computing).

29. Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. (Distributed by the author. Department of Genome Sciences, University of Washington, Seattle).

30. Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evol Bioinform Online *1*, 47-50.

Figure Legends

Figure 1. Multidimensional scaling plot of European Romani, non-Romani European and Indian individuals (see Figure S4 for more details on Indian and non-Romani European individual samples).

Figure 2. Neighbor-Joining tree based on Weir and Cockerham´s F_{ST} distances among European Romani, non-Romani European and Indian populations.
The thickest branches have at least 95% bootstrap support, and the branches of intermediate thickness have at least 75% support in 1,000 bootstrap samplings. According to AMOVA analysis, 2.71% of the genetic variation was present within the Romani (2.16% without considering the Welsh Romani), 0.47% within non-Romani European and 2.42% (1.54% without Chenchu) within Indian groups (all p-values <0.0005).

Figure 3. Population structure of European Romani, non-Romani European and Indian individuals as inferred by ADMIXTURE.
Each vertical bar represents an individual and the proportion of each individual to the *k* ancestral components is shown in colours. Linguistic classification of the Indian populations is (in same order as in the figure): Austro-Asiatic (Santhal and Kharia), Indo-European (from Kashmiri Pandit to Satnami) and Dravidian (from Chenchu to Kurumba).

Figure 4. Reconstructed demographic history of the European Romani.
The width of the branches is proportional to the estimated effective population sizes (East Asia 2417 chromosomes, India 3028, proto-Romani 1680, Spanish Romani 1090, Bulgarian Romani 1056, and non-Romani Europeans 2965). Arrow width indicates migration rates (non-Romani Europe to India 0.0003, East Asia to India 0.0004, non-Romani to Bulgarian Romani 0.036, non-Romani to Spanish Romani 0.04, in units of fraction of migrant chromosomes from the donor population per generation). Time of the demographic events estimated using a generation time of 25 years.

Figure 1

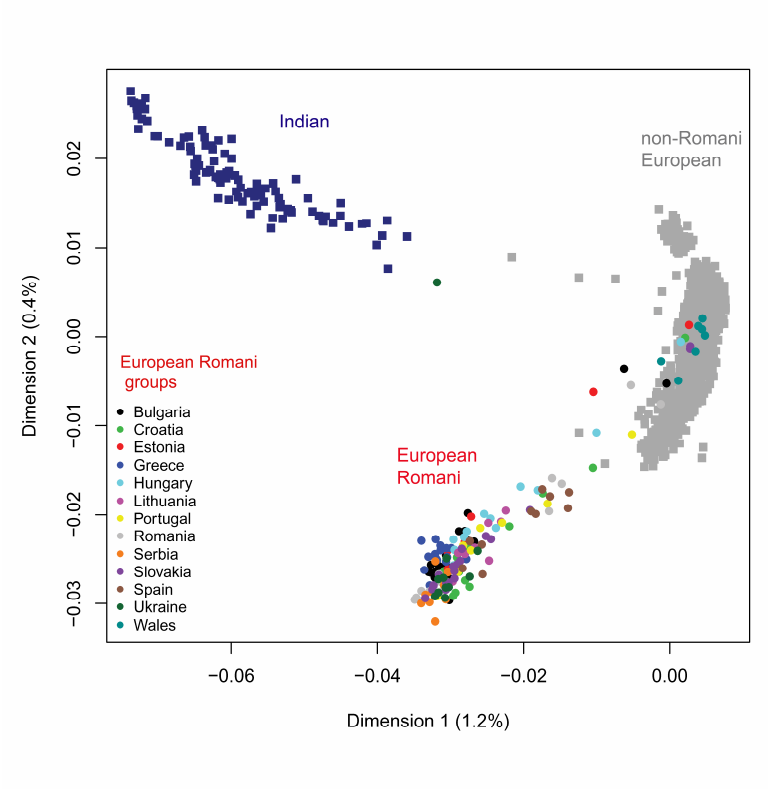

Indian
non-Romani
European
European Romani
groups
Bulgaria
Croatia
Estonia
Greece
Hungary
Lithuania
Portugal
Romania
Serbia
Slovakia
Spain
Ukraine
Wales
European
Romani
Dimension 2 (0.4%)
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.06
−0.04
−0.02
0.00
Dimension 1 (1.2%)

Figure 2

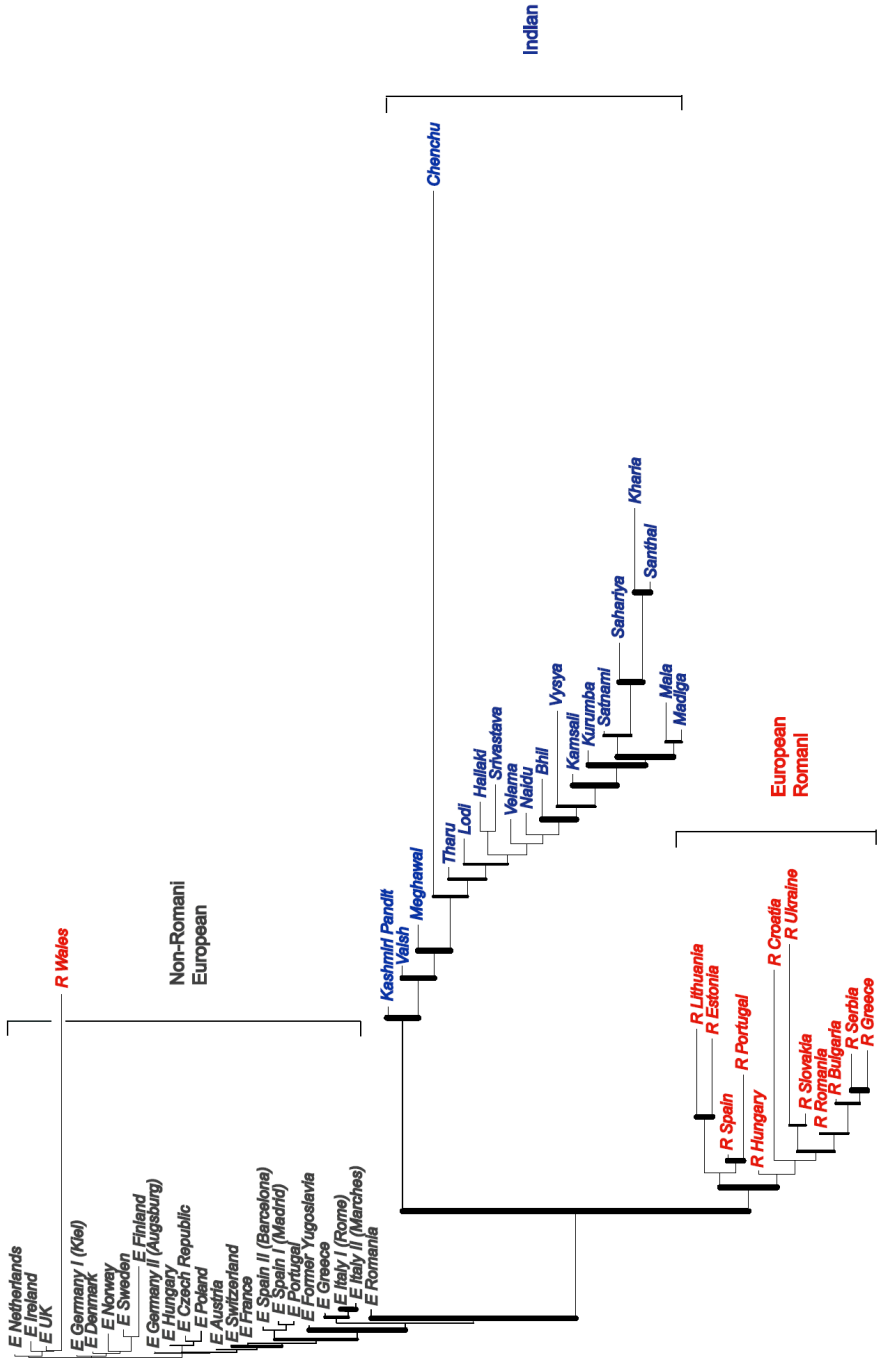

Indian
Chenchu
Kashmirl Pandit
Valsh
Meghawal
Tharu
Lodi
Hallaki
Srivastava
Velama
Naidu
Bhil
Kamsali
Kurumba
Satnami
Vysya
Saharlya
Santhal
Kharia
Mala
Madiga
European
Romanl
R Lithuania
R Estonia
R Spain
R Portugal
R Hungary
R Croatia
R Ukraine
R Slovakia
R Romania
R Bulgaria
R Serbia
R Greece
Non-Romani
European
R Wales
E Netherlands
E Ireland
E UK
E Germany I (Kiel)
E Denmark
E Norway
E Sweden
E Finland
E Germany II (Augsburg)
E Hungary
E Czech Republic
E Poland
E Austria
E Switzerland
E France
E Spain II (Barcelona)
E Spain I (Madrid)
E Portugal
E Former Yugoslavia
E Greece
E Italy I (Rome)
E Italy II (Marches)
E Romania

Figure 3

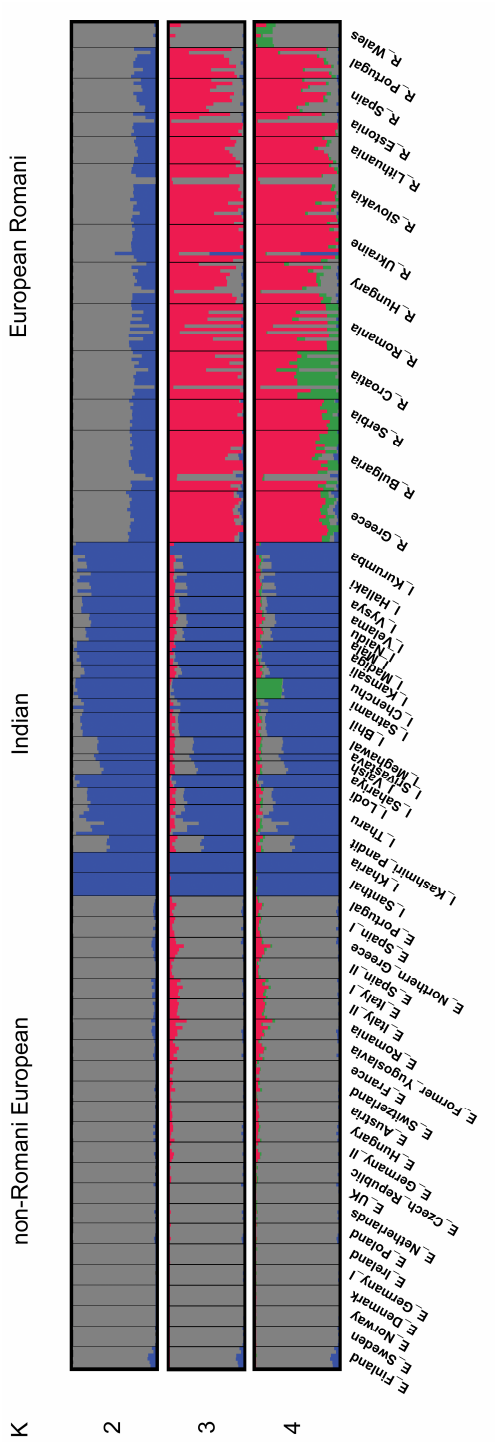
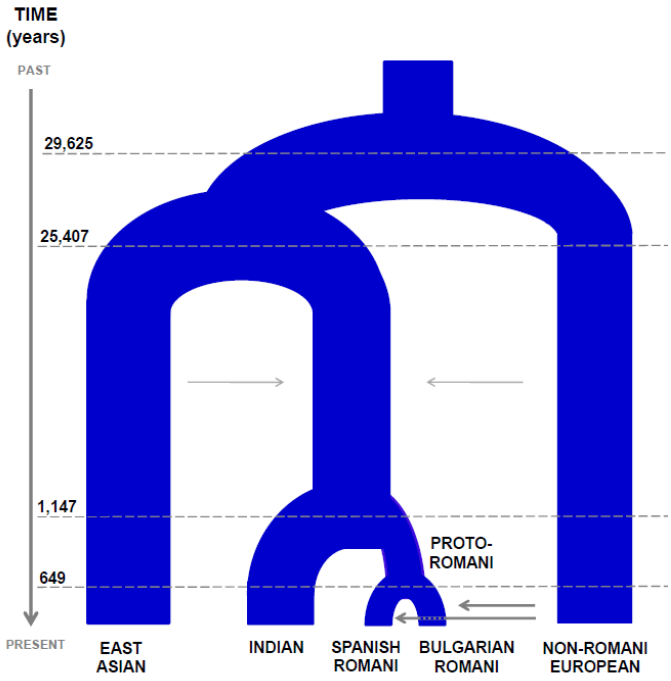

European Romani
Indian
non-Romani European
K
2
3
4
R_Wales
R_Portugal
R_Spain
R_Estonia
R_Lithuania
R_Slovakia
R_Ukraine
R_Hungary
R_Romania
R_Croatia
R_Serbia
R_Bulgaria
R_Greece
I_Kurumba
I_Hallaki
I_Vysya
I_Velama
I_Vaniya
I_Naiaa
I_Mudaliar
I_Kamsali
I_Chenchu
I_Sahnami
I_Bhil
I_Meghanwal
I_Siyasiya
I_Vaishya
I_Sanatriya
I_Lodi
I_Tharu
I_Kasshmiri_Pandit
I_Khatra
I_Santhal
E_Portugal
E_Spain_I
E_Northern_Greece
E_Spain_II
E_Italy_I
E_Italy_II
E_Romania
E_Former_Yugoslavia
E_France
E_Switzerland
E_Austria
E_Hungary
E_Germany_II
E_Czech_Republic
E_UK
E_Netherlands
E_Poland
E_Ireland
E_Germany_I
E_Denmark
E_Norway
E_Sweden
E_Finland

Figure 4


TIME
(years)
PAST
29,625
25,407
1,147
649
PRESENT
PROTO-
ROMANI
EAST
ASIAN
INDIAN
SPANISH
ROMANI
BULGARIAN
ROMANI
NON-ROMANI
EUROPEAN

3.4. Adaptive evolution of loci covarying with the human Pygmy phenotype

Isabel Mendizabal, Urko M. Marigorta, Oscar Lao, and David Comas

Human Genetics, published online 11 March 2012

4. Discussion

In this section, I will discuss the findings of this thesis and interpret them in terms of the inference of demography and adaptation from genomic data. The first sections will focus on the new insights into the demographic history of the Cuban (section 4.1.1) and Romani people (section 4.1.2). Next, lessons learnt from the methods applied on demographic inference will be discussed (section 4.1.3). Finally, the results on the adaptive history of the Pygmy phenotype will be interpreted within the context of the past and future of selection scans (section 4.2).

4.1. Demographic history

4.1.1. Cuban demographic history

Our analyses on Cuban mtDNA provide evidence that genes from extinct Native Cubans are currently segregating in the population in much higher frequency than previously thought. Our study provides a rough estimate of ~16.5% of genome-wide Native ancestry in Cubans (average of 0 and 33% in Y-chromosome and mtDNA respectively), that contrasts with previous findings. A previous study detected a maximum of 4% of native mitochondria in Cuba (Torroni et al. 1995), but it was focused on a single province in the west of Cuba. Alegre et al. (2007) analyzed the HLA locus in 68 Cubans and estimated <5% of Amerindian alleles. However, this estimate could be less reliable than ours considering the functional role of HLA in the immune system and the important role of infectious diseases in the decimation of Native Cubans. Therefore, our estimates on the amount of native contribution to current Cuban gene pool seem the most trustful ones so far, and contrast with the previous believes that Amerindian genetic ancestry in Cuba is negligible.

This result is *per se* interesting as it allows deepening in the population history of the Cuban people, but our finding also has medical implications.

Controlling for possible spurious associations in GWAS conducted on heterogeneous groups such as the Latinos in the USA is of vital importance (Choudhry et al. 2006). In addition, as observed in studies of asthma pharmacodynamics (Corvol et al. 2009), gene interactions may depend on the ancestry of the population. Any future genetic study conducted on individuals of Cuban ancestry should take this possible source of stratification into consideration.

It was also very interesting to confirm genetically that men and women contributed differentially regarding ancestry. As previously observed in other regions in America (Batista dos Santos et al. 1999; Carvajal-Carmona et al. 2000) genetics has been useful to learn from the social and political implications in mating patterns in the colonial Cuba. Whereas the Amerindian and African ancestries observed in the current population were inherited mostly from women, the contribution of the European ancestry was much larger in the paternal line.

We used the control region of the mtDNA to contrast the geographic origin of the African slaves and European settlers inferred from genetics to that from historical records. We identified Spain, and especially the Canary Islands, as the main European contributor to the current Cuban gene pool and West Africa as the main source of African slaves brought to the island. Unfortunately, the phylogenetic resolution in the HVS-I was insufficient to locate the origin of the Native Amerindian lineages in the American continent. Shortly, we will be able to learn more on the genome of one of the Caribbean Native population in Columbus' times. The Taino Genome Project (which is currently ongoing within the 1,000 Genomes Project), aims to reconstruct the genomic features of this population from 70 modern Puerto Ricans (Bustamante et al. 2011). This study may provide definitive clues on the origin of the Taino peoples in the American continent.

4.1.2. Demographic inferences on European Romani

North-West Indian origins

For the first time, we compile data that representatively covers the Indian subcontinent to test the origin of the proto-Romani population. We first approach this by means of the phylogeographic information of lineages belonging to M haplogroup of the mtDNA (section 3.2). The advantage of this approach is that in absence of recombination, and knowing that the M haplogroup is of Indian origin, the M lineages (if with enough phylogeographic resolution) should point at their geographical origin in India. Of course, the limitation lies in the assumption that the origin of these lineages is pointing at the origin of the whole Romani population.

Next, we used genome-wide SNP data (section 3.3) to locate the origin of the Romani within the Indian subcontinent (Figure 17A). Yet, it is important to account for the genetic structure of Indian populations. Specifically, Reich et al. (2009) showed that Indian populations could be accurately modelled as mixtures of two ancestral populations: Ancestral North Indian ancestry ANI (representing the West-Eurasian ancestry present in Europe, Middle East and Central Asia) and Ancestral South Indian ancestry ASI (genetically between ANI and East Asians). Accordingly, most Indian populations showed a west-Eurasian component in our analyses (see ADMIXTURE plot in section 3.3). Therefore, the recent non-Romani European admixture present in Romani populations could bias the Indian origin analysis, pointing at the Indian population with highest ANI ancestry as the source population. Consequently, we included migration from Europeans (CEU) to Indians prior to the split of the Romani in the model to test the origin of the Romani in India (Figure 17B).

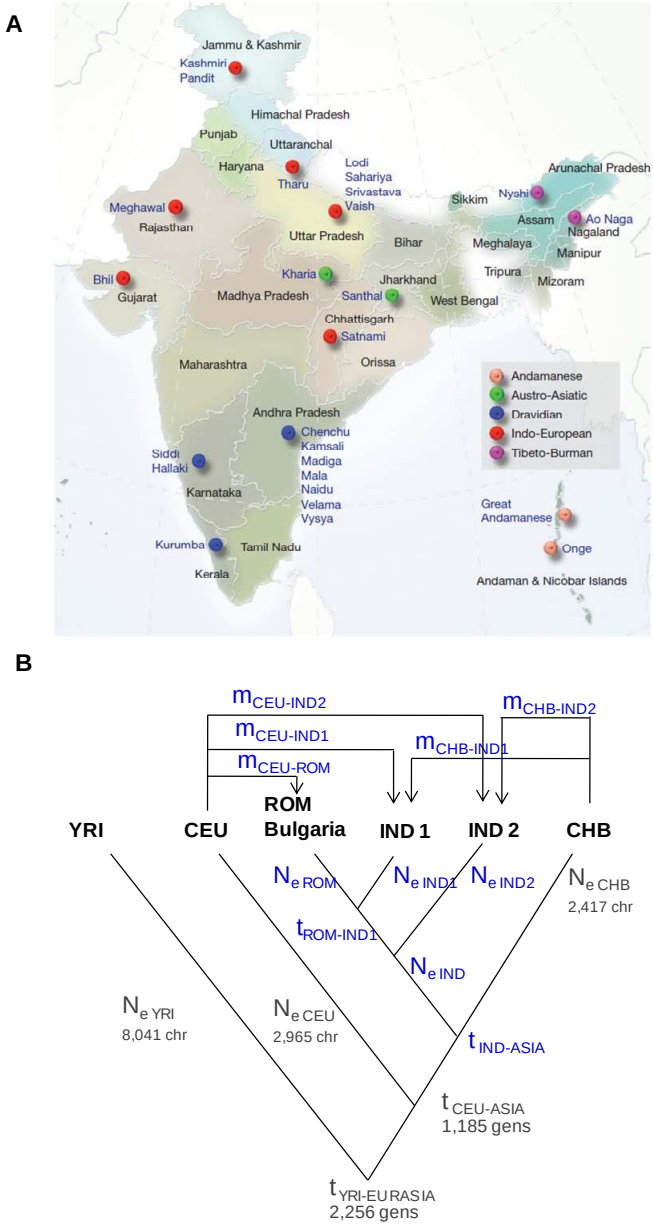


Figure 17. Samples and model used in the ABC on Romani origins. A. Map of India showing the state of origin and linguistic affiliation of the 25 groups included in our study (originally from Reich et al. (2009)). **B.** Schematic representation of the model used in ABC to estimate the Indian parental population in the Romani. Fixed parameters (and values) are shown in gray; the unknown parameters are shown in blue. This model explicitly considers migration from Europe to India prior to the split of the Romani to avoid that recent admixture between European hosts and Romani would bias the results.

Implementing complex models is not straightforward (see section 4.1.3). Still, we obtained consistent results with both approaches (mtDNA and whole-genome data). These confirmed the linguistic theories supporting a North-West Indian origin of the European Romani (most probably an Indo-European speaker population around Punjab and Rajasthan states). The use of genome-wide data permitted to date the founding event around 1.1 kya. This estimate was within the wider dates obtained from the dating the M5a1 lineage in the mtDNA.

Nevertheless, a more exhaustive representation of geography, linguistic and social strata in North-West India may be critical to exactly point at the current Indian population that is more genetically similar to the ancestral proto-Romani population. Future studies will show if the Romani originated from populations with similar anthropological features such as the Jat (Hancock 1987), as suggested by a recent study that found the same mutation causing primary congenital glaucoma (PCG) in the Punjabi Jat and European Romani families (Ali et al. 2009).

Genetic differentiation of the European Romani

Despite the relatively recent split from the parental Indian population, the European Romani show large genetic differentiation compared to current Indians. The mtDNA sequence data suggested lower maternal effective population sizes in the Romani versus non-Romani European and Indians. With genome-wide data we could quantify these differences. The founding bottleneck event in the proto-Romani group as well as the secondary bottlenecks that occurred during the settlement events in Europe (~0.7 kya) explain the severe reduction of the N_e observed in the European Romani groups of around 1/3 compared to the parental Indian population. The small effective population sizes, together with genetic isolation as testified by the long runs of homozygosity, explain the important genetic differentiation of

the Romani versus Indians, as well as the large genetic distance between European Romani groups.

The observation that the Romani genomic affinities increase with the geographic distance from the Balkan area is compatible with an initial colonization of the European continent from Southeast Europe, and a further expansion by subsequent series of founding events. In agreement with linguistic data (Fraser 1992), the strong genetic affinities found between geographically dispersed Romani populations (such as Iberian and Baltic Romani) are indicative of the accumulation of genetic drift in the consecutive migration waves after leaving the Balkan area.

This finding allows to predict that non-Balkan groups not yet characterized genetically are likely to represent a subset of the genetic diversity present in South-East Balkans, with strongest genetic affinities with the Romani populations from Eastern Central Europe (such as Hungarian and Slovakian Romani). Further studies are needed to describe the apportionment of genetic diversity in the different linguistic (such Vlax versus non-Vlax (Gresham et al. 2001)) or the religion groups in the Balkan area.

Admixture estimates

The amount of recent gene flow that the Romani received from their European hosts from autosomal data was estimated by means of ADMIXTURE software (Alexander et al. 2009). In K=3, the three ancestral components were identified as Indian, non-Romani European and Romani ancestral populations. In the case of the mtDNA, the admixture estimates were based on the identification of Romani founder lineages, which were defined as being frequent in the Romani and rare in hosts. Hence, those lineages not identified as founder provide a maximum-bound of admixture with hosts (see Figure 18A, in gray). In fact, the procedure followed to

identify founder lineages could overestimate the degree of admixture in the Balkans, as Balkan-specific lineages could be lost in other Romani populations. The autosomal data confirmed that the admixture estimates based on mtDNA were more accurate out of the Balkans (Spain, Portugal, and Lithuania). Oppositely, autosomal estimates of admixture in Bulgaria and Hungary confirmed the estimates from mtDNA as theoretical maximum bound, and suggested these could be actually much lower (see Figure 18B). Interestingly, these results suggest that the autosomal and maternal demographic histories are not divergent.

Autosomal data revealed an interesting pattern regarding gene flow with European hosts. The Romani groups that migrated out of the Balkans present more gene flow in their genomes, but in the form of short genomic fragments reflecting old admixture events. In contrast, several Balkan populations who in average have been historically stricter in endogamy rules present ongoing admixture, as testified by long chromosomal segments of European origin. This result indicates that during the expansion throughout Europe the Romani relaxed their endogamy practices during a short period, and that the current marriage rules can change from one Romani group to another.

Finally, we discarded extensive gene flow with populations geographically located in the proposed migratory path of the proto-Romani population towards the European continent. This result is in agreement with the poor influence of the Arabic in the Romani language, which locates the Romani migration before the Islamic conquests in the Middle East. Nevertheless, our results are in odds with the substantial proportion of Iranian (Kurdish and Persian) and Armenian loan words in the European Romani language (Fraser 1992). This result may indicate either that the sampling coverage of the

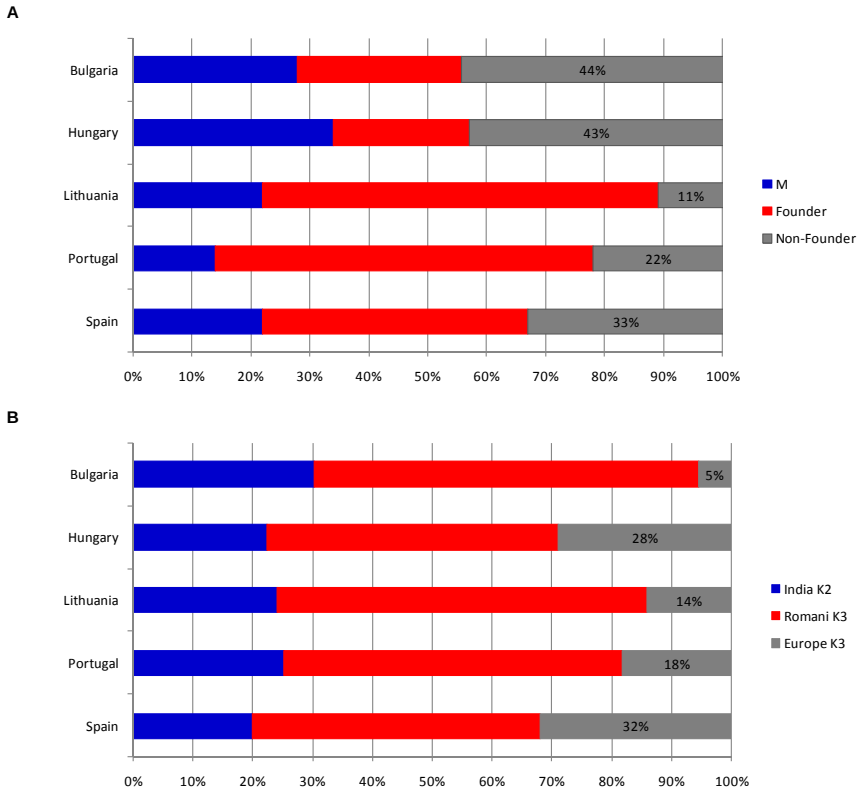


Figure 18. Admixture estimates obtained for European Romani populations that were analyzed for both mtDNA (A) and autosomal data (B). The mtDNA estimates were obtained based on the identification of founder Romani lineages based on the HVS-I (section 3.2). Admixture estimates for the whole-genome data were obtained by ADMIXTURE software (Alexander et al. 2009) (section 3.3). Although the sampling differed between the two studies, the Indian founder lineages (M in dark blue) in the mtDNA should indicate an approximated minimum threshold for the Indian ancestry in autosomal data in K=2 (dark blue). The percentage of non-founder lineages in the mtDNA (in gray) should indicate an approximated maximum limit to the percentage of recent non-Romani European ancestry in genome-wide data (K=3, in gray). As expected on how the founder lineages were identified, the admixture estimates from mtDNA between Romani and hosts should be more accurate (more similar to genome-wide data) with distance from the Balkans (see text).

actual migration route is incomplete in our study, or that the Romani received important cultural but not genetic influx from the populations during the stays in those regions. For instance, some linguists believe that the influences of Iranian and Armenian languages in the Romani could have happened already in Anatolia, since these languages (but not Arabic) were spoken in the Byzantine Empire (Matras 2002). This hypothesis would reconcile the linguistic evidence, the lack of historical records testifying a long presence of the Romani in the Middle East, and the genetic data we provide. Together, these evidence would favour a rapid migration from India to the Byzantine Empire rather than a gradual migration from India to Byzantium (Matras 2002).

Implications for complex and Mendelian diseases

The relatively high frequencies at disease mutations are *a priori* not easily explainable invoking only negative selection (Nielsen et al. 2007). Previous genetic studies on Mendelian diseases showed that the Romani founder event and endogamy practices explain the finding of the same mutations (most previously unknown) in geographically dispersed Romani groups that cause high prevalence of Mendelian disorders (Kalaydjieva et al. 2001). Accordingly, in our population-based study we find that the European Romani present a three-fold enrichment for non-synonymous (ns) SNPs in homozygous genotypes in the frequency bins of MAF <0.05 compared to the two parental groups. This excess of potentially damaging alleles suits with the expectations from a population that underwent serial bottleneck events, where purifying selection is less effective in removing these alleles (Lohmueller et al. 2008; Ohta 1973). In addition, considering the inferred past and observed current population size of the Romani (around 11 million people), this recent expansion may also have increased the non-synonymous excess in the Romani genomes (due to the genetic code, most new SNPs will

be ns), which purifying selection may not have had enough time to adjust (Lohmueller et al. 2008).

Notably, our finding does not only apply to Mendelian disorders but can also be extended to complex diseases. We found that the Romani show higher frequency at SNPs associated to metabolic/cardiovascular diseases that also present larger prevalence in the Romani. The bottleneck events, as previously shown, seem a reasonable factor for explaining prevalence differences between Europeans and Romani. Nevertheless, Indians also show higher prevalence of these diseases (which does not seem to be explained by strong bottleneck events in Indians), and some of the genes involved in metabolic disorders show signs of positive selection in Indians (Metspalu et al. 2012). These pieces of evidence suggest the thrifty gene hypothesis (Neel 1962) could hold for metabolic diseases in European Romani. However, caution is needed as the bulk of missing heritability remains yet to be discovered and future GWAS may unravel other disease variants in individuals of non-European ancestry.

The results shown in this thesis constitute the most comprehensive survey so far on the genetic characterization of the European Romani. These results have implications in anthropology, forensics, ancestry testing, and medical genomics. For instance, we noticed the presence of Romani ancestry in some European individuals that self-identified as non-Romani (see MDS plot in section 3.3), which could be a potential source of stratification in association studies done in countries with high Romani census. Finally, and similar to other founder populations, the European public health policies can largely benefit from the genomic characterization of the population history of the Romani.

4.1.3. Lessons about methods in demographic inference

Role of uniparental markers

The putative action of selection on mtDNA and Y-chromosome could invalidate certain inferences depending on the time and geographical scale considered. For instance, selection could challenge estimates of age or size of the populations (Balloux 2009). In this section, I will discuss the validity of the inferences drawn from these markers in this thesis.

The phylogeography provides a very powerful tool to do very fast and unambiguous inferences on the ancestry of a sample. For instance, in section 3.1 the uniparental markers allowed the identification of the continental ancestries in Cuba just identifying the diagnostic mutations with strong phylogeographic information, without the need of any further analyses. Another important use of the phylogeography of mtDNA was in section 3.3 where we analyzed genome-wide SNP data in the Romani. Despite having interrogated almost a million of markers, we could not be certain about the Romani origin of the Welsh samples. Since their genomes showed extensive European ancestry, we could not distinguish between the following two scenarios: i) the Welsh Romani admixed extensively with a European host population, ii) they were of European ancestry and acquired a nomadic lifestyle as in the case of native Travellers. The identification of mtDNA lineages of Indian origin (M lineages) unambiguously identified that the maternal ancestors of some Welsh Romani came from India.

Other inferences done in this thesis could be questionable. For instance, we inferred differences in population size comparing mtDNA diversity in Romani vs. non-Romani Europeans, but a non-neutral scenario could bias the estimates. However, the finding of decreased effective population size is expected in a founder population, whereas differential selection in the last millennium between European Romani and non-Romani mtDNA genes is

less plausible. Another delicate matter in mtDNA analyses is to assume the TMRCA estimates of single lineages as the age of foundation of the entire population. However, we observed that the diversification of the M5a1 lineages is private to the Romani, and that it shows a star-like network. Thus, we assumed that the timing of their diversity may provide an upper limit of the origin of the proto-Romani population. The ages obtained by dating the mtDNA M5a1 lineages match with the estimates based on the autosomal ABC as well as with historical records.

Contrasting the phylogenies of human mtDNA and Y-chromosome to those obtained from the rest of the genome seems an imperative step to definitely assess the role of selection on these markers. In any case, acknowledging that these two uniparental loci are insufficient to infer the complete demographic history of a population does not mean these are dispensable. The study of mtDNA and Y-chromosome will continue being valuable to infer demographic histories in the current context if the advantages and disadvantages of their study are reasonably weighted up.

Inferences from genome-wide SNP data

Obviously, obtaining the genotypes of several thousands of unlinked markers provides a much powerful picture of the demographic history of a population. However, the studies based on commercial SNP arrays are not exempt from important limitations. One of the biggest challenges faced in inferring demographic parameters for the European Romani (section 3.3) and African Pygmies (section 3.4.) has been the SNP ascertainment bias. This was especially critical in the case of the Romani, since the observed genetic differentiation between the parental populations (non-Romani Europeans and Indians) is low. Although we could recover most parameters with acceptable accuracy, the estimation of parameters that rely more on rare

variants (such as migration rates) can be limited by the effect of ascertainment bias.

In addition, ascertainment bias could also affect the identification of Indian and European ancestries in the Romani genomes (which we tried by different algorithms such as ADMIXTURE (Alexander et al. 2009) and HAPMIX (Price et al. 2009)). If we consider that the SNPs genotyped in commercial chips maximize the diversity within non-Romani Europeans, the power to detect European ancestry should be increased against the Indian one. Similarly, within the genomic regions of Indian origin, we could be more empowered to detect those from Ancestral North Indian (ANI) segments than Southern (ASI) ones. If we additionally take into account that the Indian segments are expected to be shorter (the last mating between Indians and Romani occurred >1.1 kya), it may explain why the identification of Indian fragments from these SNP set could be more difficult.

Interesting efforts have been made to provide a panel of SNPs with clearly documented ascertainment scheme. For instance, the Harvard HGDP-CEPH panel (http://www.cephb.fr/en/hgdp/) is a set of >600K SNPs commercialized recently by Affymetrix that is designed for demographic and selections studies, allowing also the comparison with archaic hominins and apes. Considering the decreasing costs of genotyping, the possibility to study population history with the same accuracy as it would come from deep sequencing with perfect readout of alleles (Lu et al. 2011) is very appealing and should be preferable in similar studies in the future.

The implementation and testing of complex demographic models by approximate Bayesian computation is far for being 'as simple as ABC'. Depending on the informativeness of the genetic data and the complexity of the model, model testing and parameter estimation by ABC can become a

tedious process of trial and error. Usually, after implementing the model, the user may revise it after testing the fit of the data to the model. This may imply adding or eliminating summary statistics or enclosing the range of the prior distributions. The steps to obtain the posterior distributions may also imply different options. This includes the type of distance between the simulated and the observed data, and the ABC algorithm chosen. And last, but not least, models need to be oversimplications of reality to be tractable but sufficiently complex to generate good fit to the data. Of course, comparing different models and obtaining good fit to the data does not preclude the possibility that others not included could present a better fit.

Nonetheless, ABC methods provide a flexible and powerful framework to infer demographic histories with other applications, such as the study of the molecular signatures of positive selection as shown in section 3.4. ABC algorithms are likely to be improved in the near future and may adapt well to the increasing quantity and complexity of molecular data (Csillery et al. 2010; Marjoram and Tavare 2006). An interesting methodological contribution of the work presented in this thesis is the use of the distance between multidimensional scaling plots as a summary statistic in ABC studies. Given the demographic information that these plots can provide on the population history of the samples (McVean 2009; Novembre and Stephens 2008), this statistic could outperform others that involve more drastic loss of information.

4.2. Positive selection scans, phenotypes and adaptations

Searching the genetic basis of Pygmy height

The first studies of genetic adaptations such as the HbS variants and malaria resistance (Allison 1954) were based on the comparison of the distribution of variants on candidate genes and the distribution of a phenotype or an environmental variable hypothesized to mirror a selective pressure. However, during the last years, most surveys have based only on genetic evidence and ignored the phenotypic or the environmental factors associated to adaptations. This was mainly due to two reasons. Firstly, the collection of phenotypic data is generally expensive and time consuming. Secondly, the development of high throughput genotyping technologies permitted the scan of signatures of adaptation in the whole genome without the need of *a priori* hypothesis.

These studies have allowed important development of methods to capture molecular signatures of positive selection. However, if the phenotypic or environmental factors are not included in the analyses, it is challenging to pose hypotheses on the functional target of selection. If the function of the candidate gene or region is not well known (which is usually the case) the distinction between actually selected signals and false positives (regions that show extreme value of the statistic considered due to drift effects) can be difficult.

In section 3.4. we propose the Conditional I_n statistic as a new approach for searching the genetic basis of differentiated phenotypes. The method has one main assumption: the genomic regions responsible for a population-differentiated trait must show population association with that phenotype. Being ours a population-based approach with low sample sizes, only phenotypes with a limited set of causal variants with moderate to large phenotypic effects could be detected. Nevertheless, this could be the

scenario for many adaptive phenotypes. The adaptive scenario seems plausible for Pygmy phenotype given the large population differences in height observed between Pygmy and non-Pygmy populations (30 cm in average between a Maasai and a Mbuti Pygmy) and the probable role of positive selection in the evolution of the phenotype.

The incorporation of the phenotypic (or environmental) information in population genetic analyses aimed to study positive selection is highly desirable. First, the incorporation of this piece of information is expected to reduce the number of false positives in the candidate list. As more populations with divergent demographic histories are considered in the analysis, the number of SNPs that show a distribution similar to that of the phenotype without any causality in the trait is expected to be smaller. Second, the additional information provided by the phenotype helps in the functional interpretation of the results. For instance, it improves posterior analyses conducted on the candidate list, such as functional enrichment or pathway analysis. This was observed in our results, as the pathways identified by the two statistics one considering only genetic differentiation, the other considering also phenotypic covariation, were different. Interestingly, the pathways identified by incorporating the phenotypic information made more *a priori* biological sense.

However, a definitive distinction between true and false positives is impossible only with population genetic evidence. Further studies on a model system or *in vitro* cell cultures may be necessary to provider further evidence that variants at these regions influence height through bone homeostasis, as we suggest. Alternatively, new studies based on resequencing data and larger sample sizes may help confirming or discarding the role of these regions in the Pygmy stature as well as annotating the exact causal variants.

Phenotypes and adaptations

"If all you have is a hammer, everything looks like a nail"
Abraham Maslow, 1966. *The psychology of science.*

In these years in which we are learning much about the molecular signatures of positive selection, it seems that we unconsciously try to justify all trait and functional differences by means of selective arguments. We may need to recover Gould and Lewontin's (1979) idea that functional differences are not necessarily adaptive. Body-parts, physiological systems, or molecular pathways are not isolated parts that can evolve independently. Thus, traits are under functional constraints, and the fitness of an individual will depend on the trade-off between the improvement of one trait and the simultaneous compromise on other traits.

For instance, the Pygmy stature has classically been interpreted as adaptive (i.e. the rainforest-adaptation hypotheses). However, as suggested by Migliano et al. (2007), the Pygmy low stature may not be adaptive itself. Instead, Pygmies could have strong selective pressures for an early onset of reproduction because of high adult mortality rates, being the short stature a (undesirable) consequence of the adaptation for an earlier puberty. Certainly, it is difficult to justify that early growth cessation could be beneficial itself, as according to the life-history theory, larger adult sizes imply higher fertility and reduced offspring mortality (Charnov 1993). This invokes for a need to a compensatory force. The authors show that the Pygmy fertility curves are shifted towards earlier ages, which within a context of high adult mortality rates, would offset the negative effects of short stature (see Figure 18).

.

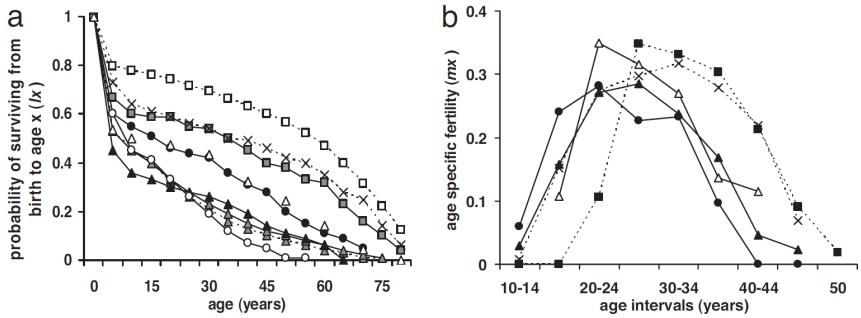


Figure 18. Age-specific survivorship (*a*) and age-specific fertility (*b*) for different human Pygmies. Pygmy populations: Eastern African Pygmies (gray triangles, dashed line), Western African Pygmies (gray triangles, solid lines), Aeta (black triangles, solid line), Batak (black circles, solid line), and Agta (open triangles, solid lines). Non-pygmy populations: !Kung (gray square, solid line), Ache (cross, dashed line), Massai (open squares, dashed line), Turkana (black square, dashed line), and chimpanzees (open circles, solid lines). The geographic locations of the Pygmy populations are shown in Figure 14. From Migliano et al. (2007).

Despite being indirectly, Migliano et al.'s hypothesis also involves the role of adaptation in the emergence of the phenotype. However, could the African Pygmy phenotype have evolved under other forces rather than positive selection? In theory, the trait could be ancestral and be maintained by purifying selection (acting on height or on another pleiotropic trait). Alternatively, and against the anthropological evidence and our results shown in section 3.4, it could be a product of genetic drift with no effect on fitness. Genetic drift plays an important role in the emergence of populations (since these typically involve population bottlenecks) and could potentially explain differential distributions of tall or short alleles between populations (Visscher et al. 2010). Nevertheless, our approach based on a differentiation-based statistic could be still valid in this case, since the method does not strictly search for selective sweeps.

Concluding remarks

The objective of this work was to decipher demography and adaptation in human populations from genomic data. Thanks to the recent technical, theoretical, and methodological developments, the study of human demography is in an effusive moment, as exemplified in this thesis with the detailed reconstruction of the Cuban and European Romani population histories. Indeed, the field holds promising future prospects. The availability of the complete spectrum of allele frequencies via resequencing will provide an unprecedented resolution of the human genetic structure at micro-geographical scales. Together with the improvements of ancient DNA techniques revealing the genomes of archaic hominins as well as of humans from different times, an exciting epoch in the study of human demographic history seems guaranteed.

The biggest promise of the genomic era is to interpret the phenotypic outcome of our genomes. In this work, a new statistic that incorporates population phenotypic data into the genetic analyses is suggested to specifically target the genomic regions associated to the Pygmy height. The gathered indirect but promising functional evidence suggest that these genomic regions may underlie the phenotype. Unfortunately, a more complete functional annotation of these regions is necessary to elaborate consistent hypothesis on the evolutionary implications of the Pygmy stature. Accordingly, the finding of the genetic basis of complex phenotypes, as reflected by genome-wide association studies, is more challenging than thought a few years ago. Without functional information, most regions in candidate lists from population-based selective scans run the risk of remaining unnoticed.

The results shown in this thesis provide strong evidence that the regions associated to the Pygmy phenotype have evolved under positive selection.

However, the models on which current selection methods are based may not be accurate for unravelling the adaptive evolution of many other interesting traits. Although the community has identified this pitfall, methodological advances are necessary to tackle with polygenic and epistatic adaptation, including that from standing variation. As proposed in this work, these future methods ought to incorporate the phenotypic data in the genetic analyses. With the immediate availability of full genome sequences, a deeper characterization of human phenotypic variation seems urgent for future advances.

Bibliography

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, Kimani J, Carrington M, Middleton D, Rajalingam R, Beksac M, Marsh SG, Maiers M, Guethlein LA, Tavoularis S, Little AM, Green RE, Norman PJ, Parham P (2011) The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. Science

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12: 1805-14

Alegre R, Moscoso J, Martínez-Laso J, Martin-Villa M, Suárez J, Moreno A, Serrano-Vela JI, Vargas-Alarcon G, Pacheco R, Arnaiz-Villena A (2007) HLA genes in Cubans and the detection of Amerindian alleles. Mol Immunol 44: 2426-35

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

Ali M, McKibbin M, Booth A, Parry DA, Jain P, Riazuddin SA, Hejtmancik JF, Khan SN, Firasat S, Shires M, Gilmour DF, Towns K, Murphy AL, Azmanov D, Tournev I, Cherninkova S, Jafri H, Raashid Y, Toomes C, Craig J, Mackey DA, Kalaydjieva L, Riazuddin S, Inglehearn CF (2009) Null mutations in LTBP2 cause primary congenital glaucoma. Am J Hum Genet 84: 664-71

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363-76

Allison AC (1954) Protection afforded by sickle-cell trait against subtertian malareal infection. Br Med J 1: 290-4

Balloux F (2009) The worm in the fruit of the mitochondrial DNA tree. Heredity 104: 419-20

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340-5

Batini C, Lopes J, Behar DM, Calafell F, Jorde LB, van der Veen L, Quintana-Murci L, Spedini G, Destro-Bisol G, Comas D (2011) Insights into the demographic history of African Pygmies from complete mitochondrial genomes. Mol Biol Evol 28: 1099-110

Batista dos Santos SE, Rodrigues JD, Ribeiro-dos-Santos AK, Zago MA (1999) Differential contribution of indigenous men and women to the formation of an urban population in the Amazon region as revealed by mtDNA and Y-DNA. Am J Phys Anthropol 109: 175-80

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162: 2025-35

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74: 1111-20

Bustamante CD, Burchard EG, De la Vega FM (2011) Genomics for the world. Nature 475: 163-5

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002) A human genome diversity cell line panel. Science 296: 261-2

Carvajal-Carmona LG, Soto ID, Pineda N, Ortiz-Barrientos D, Duque C, Ospina-Duque J, McCarthy M, Montoya P, Álvarez VM, Bedoya G, Ruiz-Linares A (2000) Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. Am J Hum Genet 67: 1287-95

Cavalli-Sforza LL (1966) Population structure and human evolution. Proc R Soc Lond B Biol Sci 164: 362-79

Cavalli-Sforza LL (1986) African Pygmies. Harcourt Brace Jovanovich, Orlando, Florida

Cavalli-Sforza LL (2005) The Human Genome Diversity Project: past, present and future. Nat Rev Genet 6: 333-40

Cavalli-Sforza LL, Wilson AC, Cantor CR, Cook-Deegan RM, King MC (1991) Call for a worldwide survey of human genetic diversity: a

vanishing opportunity for the Human Genome Project. Genomics 11: 490-1

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496-502

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38: 1251-60

Corvol H, De Giacomo A, Eng C, Seibold M, Ziv E, Chapela R, Rodriguez-Santana JR, Rodriguez-Cintron W, Thyne S, Watson HG, Meade K, LeNoir M, Avila PC, Choudhry S, Burchard EG (2009) Genetic ancestry modifies pharmacogenetic gene-gene interaction for asthma. Pharmacogenet Genomics 19: 489-96

Csillery K, Blum MG, Gaggiotti OE, Francois O (2010) Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol 25: 410-8

Curtin PD (1969) The Atlantic slave trade: a census. Madison The University of Wisconsin Press

Charnov EL (1993) Life History Invariants: Some Explorations of Symmetry in Evolutionary Ecology Oxford Univ Press, Oxford

Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai HJ, Naqvi M, Phong A, Ung N, Matallana H, Avila PC, Casal J, Torres A, Nazario S, Castro R, Battle NC, Perez-Stable EJ, Kwok PY, Sheppard D, Shriver MD, Rodriguez-Cintron W, Risch N, Ziv E, Burchard EG (2006) Population stratification confounds genetic association studies among Latinos. Hum Genet 118: 652-64

Dacal-Moure R, Rivero de la Calle M (1986) Arqueología aborigen de Cuba. Editorial Gente Nueva, La Habana

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci U S A 91: 3166-70

Diamond JM (1991) Anthropology. Why are pygmies small? Nature 354: 111-2

Edwards AW (2003) Human genetic diversity: Lewontin's fallacy. Bioessays 25: 798-801

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

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-87

Fischer A, Pollack J, Thalmann O, Nickel B, Paabo S (2006) Demographic history and genetic differentiation in apes. Curr Biol 16: 1133-8

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford

Flint-Garcia SA, Thornsberry JM, Buckler ESt (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54: 357-74

Fraser A (1992) The Gypsies. Blackwell Publishers, Oxford

Fu YX, Li WH (1993) Maximum likelihood estimation of population parameters. Genetics 134: 1261-70

Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London 205: 1161

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108: 11983-8

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paabo S

(2010) A draft sequence of the Neandertal genome. Science 328: 710-22

Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, Wise C, Angelicheva D, Calafell F, Oefner PJ, Shen P, Tournev I, de Pablo R, Kucinskas V, Perez-Lezaun A, Marushiakova E, Popov V, Kalaydjieva L (2001) Origins and divergence of the Roma (gypsies). Am J Hum Genet 69: 1314-31

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD (2011) Genetic evidence for archaic admixture in Africa. Proc Natl Acad Sci U S A 108: 15123-8

Hancock I (1987) The emergence of Romani as a koine outside of India. In: Acton T (ed) Scholarship and Gypsy struggle: commitment in Romani studies. University of Hertfordshire Press, Hatfield, pp 1-13

Hancock I (2002) We are the Romani people. University of Hertfordshire Press, Hertfordshire

Hartl DL, Clark AG (1980) Principles of population genetics. Sinauer Associates, Inc Publishers, Sunderland, Massachusetts

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M (2011) Classic selective sweeps were rare in recent human evolution. Science 331: 920-4

Hudson RR (1990) Gene genealogies and the coalescent process. Oxford Surveys of Evolutionary Biology 7: 44

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-8

Iskow RC, Gokcumen O, Lee C (2012) Exploring the role of copy number variants in human adaptation. Trends Genet 28: 245-57

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998-1003

Jobling MA, Hurles M, Tyler-smith C (2004) Human evolutionary genetics: origins, peoples, and disease. Garland Science, Abingdon and New York

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

Jombart T, Pontier D, Dufour AB (2009) Genetic markers in the playground of multivariate analysis. Heredity (Edinb) 102: 330-41

Kalaydjieva L, Gresham D, Calafell F (2001) Genetic studies of the Roma (Gypsies): a review. BMC Med Genet 2: 5

Kenrick D (2007) Historical Dictionary of the Gypsies (Romanies). Scarecrow Press, Inc, Lanham, Maryland

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56-64

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624-626

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893-903

Kimura M, Ohta T (1969) The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. Genetics 61: 763-71

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

Kingman JFC (1982) On the Genealogy of Large Populations. Journal of Applied Probability 19: 17

Lalueza-Fox C, Gilbert MT (2011) Paleogenomics of archaic hominins. Curr Biol 21: R1002-9

Lalueza-Fox C, Gilbert MT, Martínez-Fuentes AJ, Calafell F, Bertranpetit J (2003) Mitochondrial DNA from pre-Columbian Ciboneys from Cuba and the prehistoric colonization of the Caribbean. Am J Phys Anthropol 121: 97-108

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4: 203-21

Lewontin RC (1972) The apportionment of human diversity. Evolutionary Biology 6: 8

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-95

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-4

Liégeois J-P (1994) Roma, Gypsies, Travellers. Council of Europe Press, Strasbourg

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J (2000) Molecular Cell Biology. 4th edition. W. H. Freeman, New York

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD (2008) Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994-7

Lu Y, Patterson N, Zhan Y, Mallick S, Reich D (2011) Technical design document for a SNP array that is optimized for population genetics.

Mainard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23-35

Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci U S A 100: 15324-8

Marjoram P, Tavare S (2006) Modern computational approaches for analysing molecular genetic variation data. Nat Rev Genet 7: 759-770

Matras Y (2002) Romani: a linguistic introduction. Cambridge University Press, Cambridge

McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genet 5: e1000686

Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, Nelis M, Magi R, Metspalu E, Remm M, Pitchappan R, Singh L, Thangaraj K, Villems R, Kivisild T (2012) Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am J Hum Genet 89: 731-44

Migliano AB, Vinicius L, Lahr MM (2007) Life history trade-offs explain the evolution of human pygmies. Proc Natl Acad Sci U S A 104: 20216-9

Neel JV (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14: 353-62

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857-68

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40: 646-9

Ohenjo N, Willis R, Jackson D, Nettleton C, Good K, Mugarura B (2006) Health of Indigenous people in Africa. Lancet 367: 1937-46

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246: 96-8

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 22: 201-4

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S, Kidd KK, Kidd JR, Van der Veen L, Hombert JM, Gessain A, Froment A, Bahuchet S, Heyer E, Quintana-Murci L (2009)

Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet 5: e1000448

Pérez de la Riva J (1979) El monto de la inmigración forzada en el siglo XIX. Editorial Ciencias Sociales, Ciudad de la Habana

Perry GH, Dominy NJ (2009) Evolution of the human pygmy phenotype. Trends Ecol Evol 24: 218-25

Perry JR, Stolk L, Franceschini N, Lunetta KL, Zhai G, McArdle PF, Smith AV, Aspelund T, Bandinelli S, Boerwinkle E, Cherkas L, Eiriksdottir G, Estrada K, Ferrucci L, Folsom AR, Garcia M, Gudnason V, Hofman A, Karasik D, Kiel DP, Launer LJ, van Meurs J, Nalls MA, Rivadeneira F, Shuldiner AR, Singleton A, Soranzo N, Tanaka T, Visser JA, Weedon MN, Wilson SG, Zhuang V, Streeten EA, Harris TB, Murray A, Spector TD, Demerath EW, Uitterlinden AG, Murabito JM (2009) Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. Nat Genet 41: 648-50

Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, Shriver MD (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68: 198-207

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5: e1000519

Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. Nat Rev Genet 11: 665-7

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20: R208-15

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-59

Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S, Balanovsky O,

Kidd KK, Kidd JR, van der Veen L, Hombert JM, Gessain A, Verdu P, Froment A, Bahuchet S, Heyer E, Dausset J, Salas A, Behar DM (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci U S A 105: 1596-601

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A 102: 15942-7

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Paabo S (2011) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053-60

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature 461: 489-94

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411: 199-204

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73: 1402-22

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1: e70

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298: 2381-5

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-7

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. Science 312: 1614-20

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-8

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-25

Sayer JA, Harcourt CS, Collins N (1992) The conservation atlas of tropical forests: Africa. IUCN, WCMC, Macmillan, London

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576-83

Shea BT, Bailey RC (1996) Allometry and adaptation of body proportions and stature in African pygmies. Am J Phys Anthropol 100: 311-40

Slatkin M (1991) Inbreeding coefficients and coalescence times. Genet Res 58: 167-75

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84: 740-59

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-95

Tattersall I (2009) Becoming Modern Homo sapiens. Evo Edu Outreach 2: 584–589

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505-18

Teo YY, Small KS, Kwiatkowski DP (2010) Methodological challenges of genome-wide association analysis in Africa. Nat Rev Genet 11: 149-60

The 1000 Genomes Project Consortium (2011) A map of human genome variation from population-scale sequencing. Nature 467: 1061-73

The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789-96

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM (2009) The genetic structure and history of Africans and African Americans. Science 324: 1035-44

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31-40

Torroni A, Brown MD, Lott MT, Newman NJ, Wallace DC (1995) African, Native American, and European mitochondrial DNAs in Cubans from Pinar del Rio Province and implications for the recent epidemic neuropathy in Cuba. Cuba Neuropathy Field Investigation Team. Hum Mutat 5: 310-7

Underhill PA, Kivisild T (2007) Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet 41: 539-64

Visscher PM, McEvoy B, Yang J (2010) From Galton to GWAS: quantitative genetics of human height. Genet Res (Camb) 92: 371-9

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72

Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF (2008) A novel DNA sequence database for analyzing human demographic history. Genome Res 18: 1354-61

Watterson GA, Guess HA (1977) Is the most frequent allele the oldest? Theor Popul Biol 11: 141-60

Weiss KM, Long JC (2009) Non-Darwinian estimation: My ancestors, my genes' ancestors. Genome Research 19: 703-710

Wilkins JF (2006) Unraveling male and female histories from human genetic data. Curr Opin Genet Dev 16: 611-7

Wright S (1931) Evolution in Mendelian populations. Genetics 16: 36

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

Wright S (1951) The genetical structure of populations. Annals Eugenics 15: 32

Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74: 50-61

Appendix

Contributions to other articles

A1. SNPlexing the human Y-chromosome: A single-assay system for major haplogroup screening

Gemma Berniell-Lee, Karla Sandoval, **Isabel Mendizabal**, Elena Bosch, and David Comas

Electrophoresis. 2007 Sep;28(18):3201-6:

A.2. Genetic structure of Tunisian ethnic groups revealed by paternal lineages

Karima Fadhlaoui-Zid, Begoña Martinez-Cruz, Houssein Khodjet-el-khil, **Isabel Mendizabal**, Amel Benammar-Elgaaied, and David Comas

American Journal of Physical Anthropology 2011 Oct;146(2):271-80

A3. Y-chromosome diversity in Native Mexicans reveals continental transition of genetic structure in the Americas

Karla Sandoval, Andrés Moreno, **Isabel Mendizabal**, Peter Underhill, María López-Valenzuela, Rosenda Peñaloza-Espinosa, Leonor Buentello-Malo, Heriberto Avelino, Francesc Calafell, and David Comas

American Journal of Physical Anthropology 2012 july; 148(3):395-5
Published online May 11, 2012