Genetic variation of the X chromosome and the genomic regions of Coagulation Factors VII and XII in human populations: Epidemiological and evolutionary considerations

Georgios Athanasiadis

Genetic variation of the X chromosome and the genomic regions of Coagulation Factors VII and XII in human populations: Epidemiological and evolutionary considerations

Doctoral thesis presented by

Georgios Athanasiadis

in solicitation of the degree of
Doctor of Philosophy awarded by the University of Barcelona

Directed by Dr. Pedro Moral Castrillo, Professor of Physical Anthropology
at the Unit of Anthropology, Department of Animal Biology
University of Barcelona

Doctorate Programme of Biodiversity
Department of Animal Biology – Faculty of Biology

Pedro Moral Castrillo
Director

Georgios Athanasiadis
Doctorate student

Variación genética en el cromosoma X y las regiones genómicas de los Factores de Coagulación VII y XII en poblaciones humanas: Consideraciones epidemiológicas y evolutivas

Tesis doctoral presentada por

Georgios Athanasiadis

para optar al grado de
Doctor por la Universidad de Barcelona

Dirigida por el Dr. Pedro Moral Castrillo, Profesor Titular de Antropología Física
de la Unidad de Antropología del Departmento de Biología Animal
de la Universidad de Barcelona

Programa de Doctorado de Biodiversidad
Departmento de Biología Animal – Facultad de Biología

Pedro Moral Castrillo
Director

Georgios Athanasiadis
Doctorando

Αφιερωμένο στη μνήμη του παππού μου,
Παύλου Δαμιανού

Dedicated to the memory of my grandfather,
Pavlos Damianos

O princípio da ciência é sabermos que ignoramos.

Fernando Pessoa, A Hora do Diabo

ACKNOWLEDGEMENTS

Barcelona, 11 April 2010

It has been more than five years since I left my life in Thessaloniki for Barcelona. Looking back, I can see in myself of those days some of the qualities that brought me where I am today: innocence, effortless optimism and decisiveness. I am so grateful that reason was not my foremost consultant back then! Had it been so, I would never have renounced the comfort of my Greek lifestyle (my compatriots should know exactly what I mean). I now am one step before finishing my doctoral studies and I feel the need to thank all the people who had something to do with it.

The first person to thank could be no other but my supervisor, Pedro Moral. With his positive answer to my first email, he opened a door of unimaginable possibilities for me. *Pedro, gracias por enseñarme cómo hacer ciencia, por nunca decirme un no, por tratarme como a un hijo…*

I would also like to thank all the people in the Moral lab: Esther, Marc, Emili, Josep and Magda R for being next to me when I first joined the group, but also the people who came later, Magda G and Robert. *Thank you for all the help you gave me and for the endless hours of conversation and exchange of ideas. You have been the best company ever and made all this worth even more!* Also, many thanks to the Moral people I met throughout these years but haven't had the opportunity to know better: Toni, Neus V, Natalia, Meri, Eva and Ares.

But my doctoral studies were not all about Barcelona; in the last three years I had the rare fortune to visit three very important research centres outside Spain, which gave a real boost to my CV. I therefore am eternally grateful to Dr. Cathryn Lewis and Dr. Mike Weale from King's College London, Dr. Mark Stoneking from the Max Planck Institute for Evolutionary Anthropology in Leipzig and Dr. John Hopper from the University of Melbourne for hosting me at their labs and making me feel like home. Also, many thanks to all the wonderful people I met in these places together with my apologies for not mentioning each one of them separately.

From this section, how could there be missing a special mention to all the people from the "Unitat d'Antropologia"? *Mireia, Neus M and Marta (the best representation of The Three Graces ever!); Bàrbara, Araceli, Mar, Mari, Nàdia, Sílvia and the rest of the Lourdes Girls (my apologies to Sergi!); Jordi, Bea, Ferràn, Laura and Mohammed; and last but not least the "big bosses" – Lourdes, Clara, Miquel, Alejandro and Txomin –, thank you for all the little things that spiced up our day-to-day routine at work! A més, moltes gràcies a Dr. Pons per la seva amabilitat i noblesa. Vostè és tota una inspiració per a mi!*

There are also so many friends who stood next to me throughout these years. Starting from the closest and oldest ones, I want to thank Christina and Panagiotis for their pure friendship, devotion and moral support. *Είμαι πολύ τυχερός που είστε φίλοι μου!* Also, many thanks to the rest of my friends – older and more recent – for all the happy moments we shared: Vasilis D, Vasilis B, Despoina, Modesta, Albert, Matteo, Serguei, Alessandro, Mario, Rossella, Magie, Oseas, John, Savvas… the list is endless!

Special thanks to my dear friend James for reading through my thesis and making all the necessary language corrections. *James, this work would not be the same without your help.*

Finally I feel the need to thank my family, my beloved brother and sister – Damon and Sophie –, my mum and dad – Chrysoula and Panagiotis – and my granny Kyriaki for their unconditional love and silent but constant moral and material support. It took me some time to realise it, but I now see it clearly: family is where everything starts and where everything ends. *Χωρίς εσάς δε θα ήμουν ούτε το μισό από αυτό που είμαι τώρα.*

Contents

Table of Contents

Introduction

GENETIC VARIATION IN HUMANS

Like most living organisms, human beings vary in almost all their characteristics. Much of this variation is found in the genome and can either have an observable manifestation or not. For example, variation in hair, eye or skin colour, as well as in blood groups, HLA type, even response to drugs and susceptibility to infectious (e.g. malaria) or chronic diseases (e.g. diabetes, cancer, schizophrenia or cardiovascular disorders) are all differences in the phenotype caused – to different extent – by differences in the genotype. However, the overwhelming majority of genetic variation plays no role in the phenotype. This fraction of genetic variation is often referred to as neutral variation because, as we shall see further down, it has no evolutionary consequences.

Genetic variation is the raw material of human population genetics. Each fraction of genetic variation reflects different aspects of human evolution. On the one hand, neutral variation, being unnoticed by natural selection, provides a passive record of human demographic history. Thus, neutral variation can for example be used to identify past population movements or episodes of growth and decline in population size. On the other hand, variation with biological and evolutionary consequences plays a key role in the understanding of another aspect of human history – the biological – and has revealed important genetic aspects of the processes that moulded our species (Harpending and Cochran, 2005).

Processes that shape genetic variation

The genetic variation we observe in the present human populations is the composite result of five processes; mutation, recombination, genetic drift, selection and migration. In brief, mutation generates new alleles, while recombination shuffles the pre-existing genetic variation, both acting on the gametes. Conversely, the other three processes act as allele frequency modifiers on the populations; new alleles can be wiped out of the gene pool or increase in frequency through stochastic processes (genetic drift), selection changes the allele frequencies of non-neutral loci by acting on the survival and reproductive ability of the individuals and, finally, migration affects the genome-wide geographic distribution of the human genetic diversity. The following paragraphs contain a more detailed presentation of these five processes as well as several relevant population genetic concepts that are used in this work.

Mutation

Mutation is any change in the DNA sequence. The term encompasses a wide range of events such as substitutions of a single base or small insertions and deletions of a few bases. More complex mutational changes include expansions or contractions in the number of tandemly repeated DNA motifs; insertions of transposable elements; insertions, deletions, duplications and inversions of megabase-long segments of DNA; translocations of chromosomal segments; and even changes in chromosomal number. Each kind of mutation is characterised by a distinct molecular mechanism and rate.

Mutations are usually caused by mutagenic factors, such as radiation, chemicals and viruses, which affect the mechanisms of DNA replication and repair in a variety of ways. Random errors during DNA replication or meiosis are also an important source of mutation.

Mutation is the only process that generates new alleles, thereby providing the raw material on which selection and other evolutionary forces can act. Although mutations occur constantly in the DNA of all cells of the body, only those that occur in the germ line can pass to the next generation and contribute to the evolutionary change. To do so, they also have to be nonlethal and compatible with fertility. Mutations in other cells of the body (somatic mutations) are not heritable but they can have serious consequences on the survival of the individual.

Recombination

Recombination is the exchange of DNA segments between aligned paternal and maternal chromosomal homologues during meiosis. This process (also known as crossover) is reciprocal in the sense that no net loss of genetic information occurs. Although recombination does not produce new variation like the mutations do, it generates new combinations of pre-existing alleles at different loci (haplotypes), therefore affecting the ability of populations to adapt to their environment.

Recombination events disrupt the coinheritance of molecular markers. Thus, they can be studied at the population level by investigating whether specific alleles at different loci are associated with one another more or less often than

would be expected by chance. This nonrandom association is known as linkage disequilibrium (LD) and has invaluable applications in human population genetics and genetic epidemiology (Figure 1).

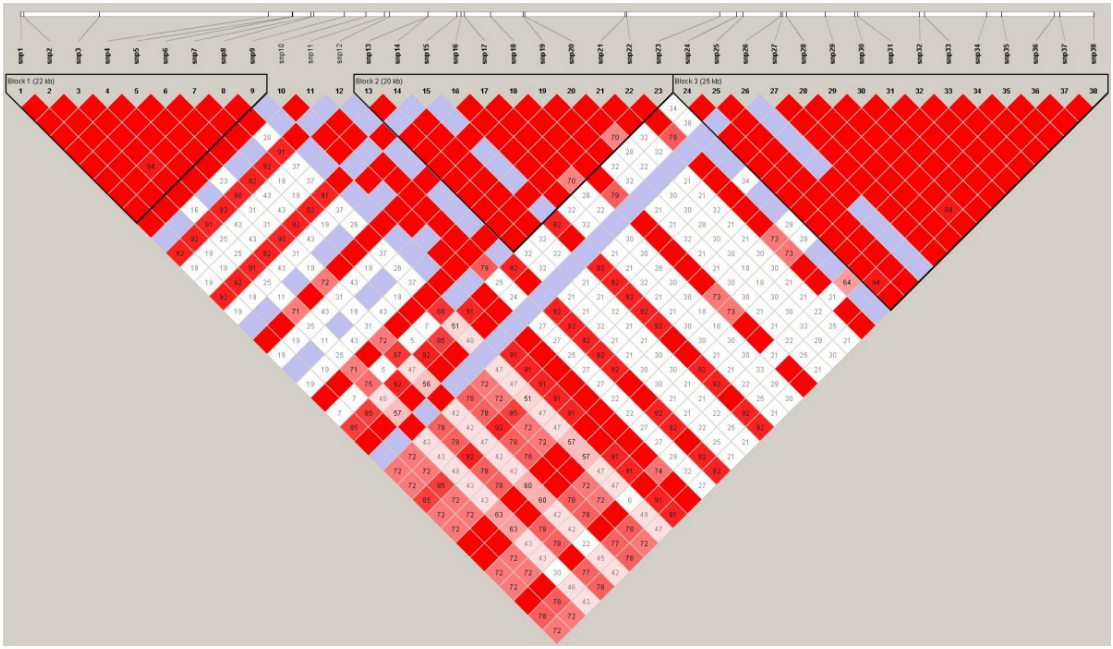


Figure 1: Linkage disequilibrium pattern of 38 single nucleotide polymorphisms from the GCH1 (GTP cyclohydrolase 1) gene in Hispanic Americans. The analysis was performed with the Haploview software and revealed three major haplotype blocks (image taken from Kim and Dionne, 2007).

The simplest model of recombination assumes a uniform rate along the entire DNA molecule. As a consequence, the probability of a crossover between a pair of markers will depend only on the physical distance that separates them. However, empirical data suggest that recombination is a more complex process. On the one hand, not all recombination events result in a crossover; in some cases the result is a nonreciprocal exchange called gene conversion. On the other hand, recombination rates were found not to be uniform along a segment of DNA, as crossovers appear to be concentrated in 'recombination hotspots' between 'cold'

regions. Besides, the distribution of recombination is known to vary between the sexes, with females undergoing more recombination events than males do (around 80 and 50 recombination events per meiosis respectively; reviewed by Jobling et al., 2004).

Genetic drift

Genetic drift is the fluctuation of allele frequencies in a finite population due to the random variation in the contribution of each individual to the next generation. Since populations are not infinitely large, each generation represents a finite sample from the previous one. As a consequence, there is a possibility that variation in allele frequency between generations is solely due to the stochastic process of sampling.

The magnitude of genetic drift is related to the size of the population sampled. Figure 2 shows the change in allele frequency over 50 generations in simulated populations of different sizes, starting with an initial allele frequency of 0.5. Fluctuations in allele frequency are dramatic in populations of size 20, with the allele in some cases rapidly becoming fixed or lost. Conversely, in populations of size 200, even though the fluctuations are still notable, no allele fixations or losses are observed. Finally, more subtle variations in frequency occur in populations of size 2000.

As real human populations present different degrees of generation overlapping, size constancy and mating randomness, a straightforward comparison of the genetic drift in different populations is hard to achieve; to bypass this difficulty, Wright's effective population size (N_e) is used instead. The effective population

size of a real population represents the size of an idealised, Wright-Fisher population (i.e. featuring no overlapping generations, constant size and random mating) that experiences the same amount of genetic drift as the one under study. N_e is a measure of the genetic drift: the smaller the N_e, the greater the drift.

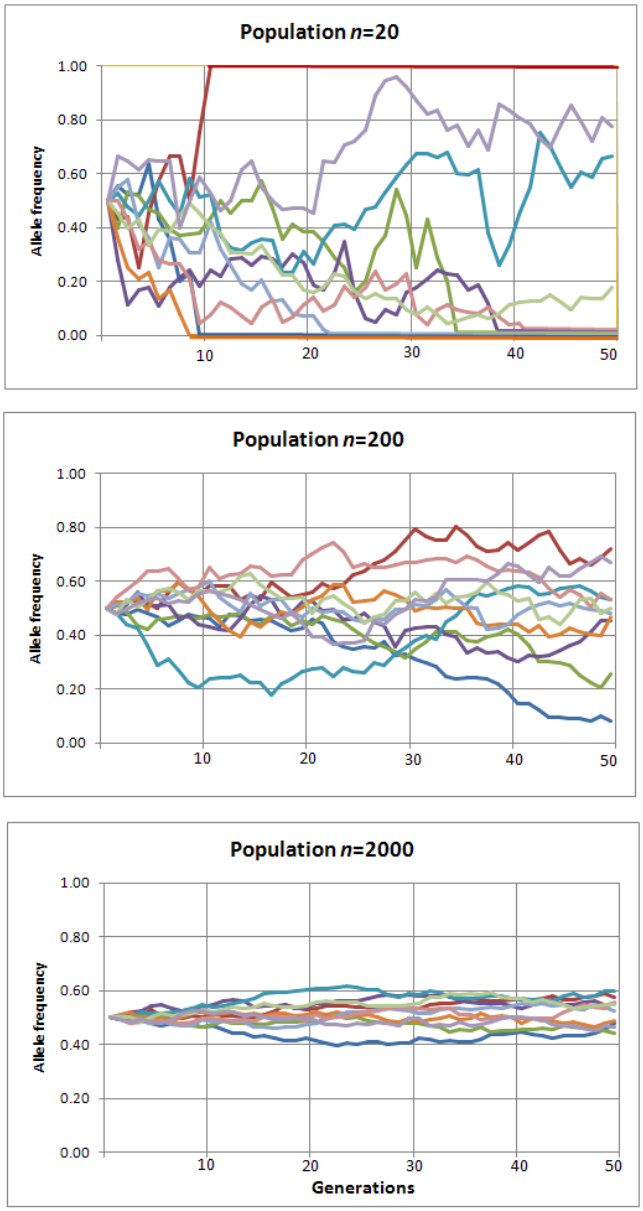


Figure 2: Ten simulations of random genetic drift of a single allele measured over 50 generations repeated in three populations of different sizes (20, 200 and 2000). In general, alleles in smaller populations drift faster to loss or fixation.

The long-term effective population size has been shown to be approximately equal to the harmonic mean rather than the arithmetic mean of the population

sizes over time (Wright 1938; Crow and Kimura 1970). This means that the N_e is disproportionately affected by small population sizes and explains why human effective population size is still very small (latest estimates based on genome-wide SNP data: ~7,500 for Yoruba from Ibadan, Nigeria; ~3,100 for the remaining three non-African samples from HapMap Phase 1; Tenesa et al., 2007).

Selection

Natural selection is a key mechanism of evolution; it can be defined as the process by which heritable traits that make it more likely for an organism to survive and successfully reproduce become more common in a population over successive generations. In other words, natural selection causes the differential reproduction of genotypes in succeeding generations. For selection to operate, these genotypes need to have a phenotypic outcome.

Selection can occur at any stage throughout the life of an individual and can involve survival into reproductive age, success in attracting a mate (sexual selection) and success in producing offspring (as reflected by fertility and fecundity). These characteristics make up the fitness of an individual – its ability to survive and reproduce. Fitness depends both on the genetic composition of an individual and its environment.

Selection can have different manifestations depending on the effect of a mutation on the phenotype. Mutations that have a positive effect on the survival and reproduction success of an individual are the most likely to be passed to the next generation. Such mutations are subject to positive selection. On the other hand,

mutations that reduce the fitness of the carrier are the most likely to be erased from the gene pool, undergoing negative selection.

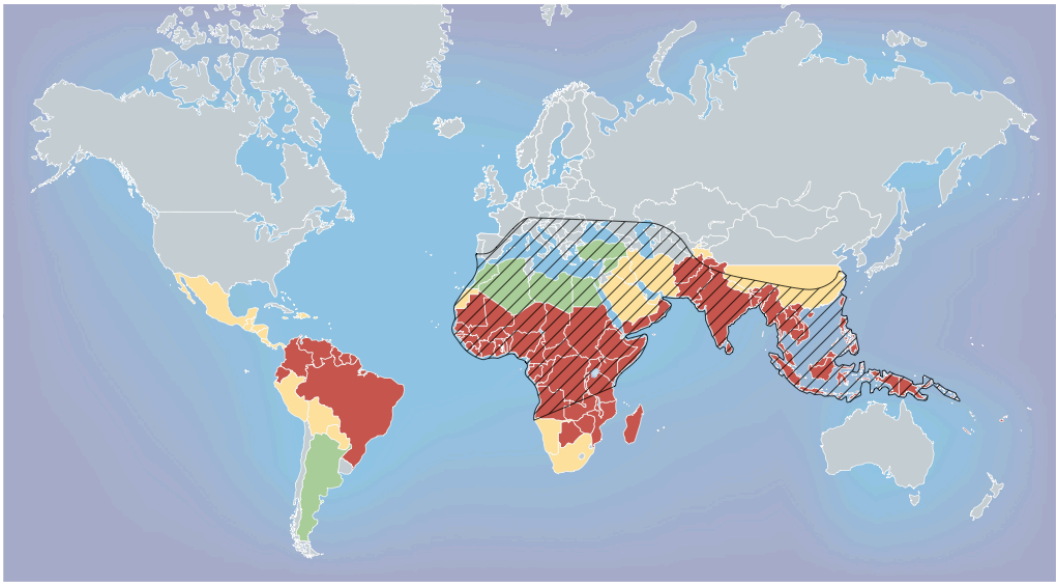


Figure 3: Global distribution of malaria and red-blood-cell disorders. Green indicates areas where malaria is only present in a few remote locations, yellow indicates areas with intermediate malaria risk and red indicates areas with high malaria risk. The hatched area shows the distribution of red-blood-cell disorders. Source of data: World66.com and Weatherall et al., 2001 (image taken from Cooke and Hill, 2001).

The two types of selection mentioned above reduce genetic diversity through either the overrepresentation of an allele in the next generation or its elimination from the gene pool. However, in some cases selection can lead to an increase of genetic diversity. This is the case of balancing selection, whereby a new allele may increase the fitness of a heterozygote because both homozygous genotypes reduce fitness. A classic example of a balanced polymorphism is the sickle cell anaemia allele Hb^S (Haldane 1949), which protects against malaria when heterozygous but dramatically reduces fitness when homozygous due to severe red blood cell disorders. As a consequence, red blood disorders caused by

polymorphisms that protect from malaria present a similar geographic distribution with malaria itself (Figure 3).

Finally, a large number of mutations are never subject to any kind of selection because they are located in that fraction of the human genome (accounting for the ~98% of the total size) that does not contain any functional genes. Such mutations are known as neutral mutations because they are not submitted to any selective pressure. Thus, their perseverance in the population depends on stochastic processes only.

Migration

Migration is the systematic movement of people and their genes from one population to another. It holds a special place in human genetic diversity due to the complex demographic history of human populations. Unlike other evolutionary processes, migration cannot change allele frequencies on a species level but is capable of changing allele frequencies within a subpopulation. Moreover, these changes affect the whole genome rather than just a specific locus. Migration is often used as a synonym for gene flow although, strictly speaking, for gene flow to occur the migrants must contribute to the next generation in their new location. Gene flow tends to reduce inter-population diversity and increase intra-population diversity.

Many different models of migration have been proposed to describe gene flow. According to Wright's island model, a meta-population is divided into 'islands' of equal size N, which exchange genes at the same rate m per generation (Figure 4). Other assumptions of this model are the lack of any other geographical

substructure apart from the division into islands, the infinite persistence of all populations and the lack of selection or mutation. The rate of migrant exchange (Nm) can be related directly to Wright's F_{ST} estimates of population structure by the equation:

$$F_{ST} = \frac{1}{1 + 4Nm}$$

Kimura and Weiss's stepping-stone model (Kimura and Weiss, 1964) is an improvement of Wright's island model, as it introduces the concept of geographic substructure. According to this model, gene exchange is only permitted between adjacent populations. However, here too migration rates between subpopulations are equal (Figure 4).

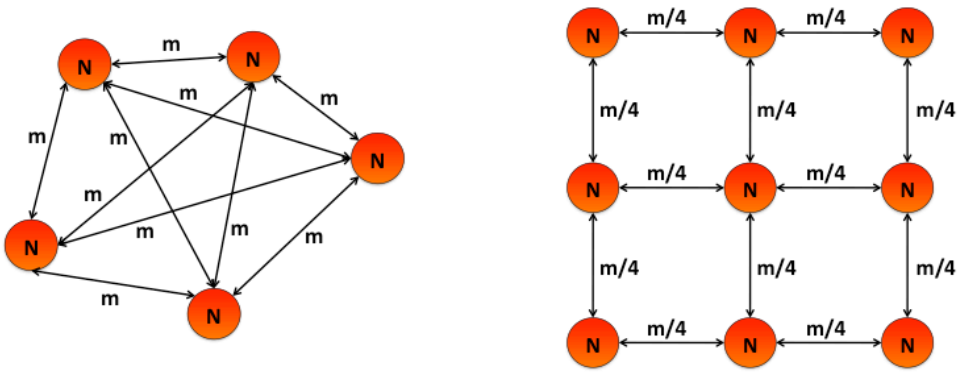


Figure 4: The island (on the left) and stepping-stone (on the right) models of gene flow. N: population size; m: rate of exchange of genes per generation (image adapted from Jobling et al., 2004)

Rather than assuming discreet subpopulations, isolation by distance models treat migration as occurring within a continuous population (Wright, 1943; Malécot, 1948) where mating choices are limited by distance. Within such models, genetic similarity develops in neighbourhoods as a function of dispersal

distances. The aforementioned stepping-stone model is a discontinuous example of isolation by distance.

Molecular markers and genetic analysis: a retrospect

Any protein system or fragment of DNA that presents variation can be used as a molecular marker. The development of molecular markers dates back to 1900, when Karl Landsteiner discovered the AB0 blood group system (Landsteiner, 1900), the first molecular marker ever to be described – an important development in the field of human population genetics. Population studies of the AB0 variation started as early as 1919 (Hirszfeld and Hirszfeld, 1919). Later, more blood groups (such as MN and RH) were described and put into use. In the following decades, immunological methods were put forward to describe molecular variation in the immunoglobulins and the extremely polymorphic HLA system. With the introduction of protein electrophoresis, more markers became available, such as haemoglobins and other blood proteins. This last method consists in distinguishing protein variants according to differences in size and charge caused by amino acid substitutions.

All the above molecular markers have one thing in common: they are to some extent gene products. One of the disadvantages of protein markers is that they are an indirect and insensitive method of detecting variation in the DNA. A more direct molecular marker would survey DNA variation itself, rather than rely on variations in the electrophoretic mobility of proteins that the DNA encodes. In this context, the advent of the undoubtedly advantageous techniques of DNA analysis caused research interest to shift from protein-based to DNA-based markers (Figure 5).

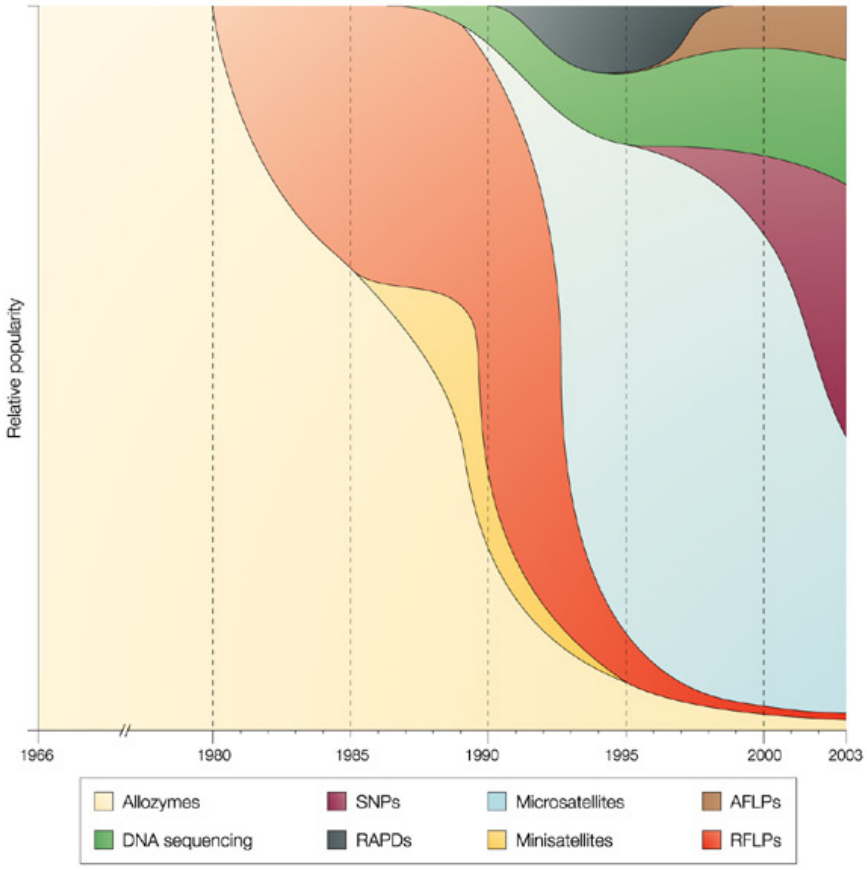


Figure 5: Subjective view of the changing relative importance of different molecular markers (taken from Schlötterer, 2004).

The first DNA markers to be used were the restriction fragment length polymorphisms (RFLPs), back in the 1960s, in which a single base substitution in the recognition sequence of a restriction enzyme changed the pattern of the resulting restriction fragments. The first association studies and genetic maps were produced thanks to these markers. However, technical limitations precluded the wider exploitation of RFLPs. In the 1970s Alu insertions were also discovered by virtue of restriction enzymes; because of their valuable properties, these markers were used in human population genetics and primate phylogenetics (see next section for more details). Restriction enzymes were also involved in another kind of polymorphism, the minisatellites, consisting of

tandem repeats that present length polymorphism. Minisatellites revolutionised the genetic identification of individuals and were successfully applied to forensics and paternity testing. However, the complexity of their banding patterns and their non-random distribution in the genome hindered their use in population genetics, genome mapping and association studies.

By the mid-1980s, the polymerase chain reaction (PCR) was invented (Saiki et al., 1985). PCR brought about a real revolution in the use of DNA molecular markers, as now any genomic region could be amplified and analysed without the requirement for cloning or isolating large amounts of ultrapure genomic DNA. The first widespread markers to be exploited by PCR technology were the microsatellites. With a typical repeat region smaller than 100 bp, most microsatellites can be amplified by a standard PCR. Microsatellites were applied successfully to genome mapping, paternity testing and population genetics (see next chapter for more details). The importance of these markers in the research can be seen in the fact that the now well-known increased genetic diversity of Africans as compared to other human groups was not apparent until ascertainment bias-free genetic markers like microsatellites became available (Harpending and Cochran, 2005).

In the 1990s yet more PCR-based markers appeared. Their common feature was the use of PCR primers that can bind to multiple sites in the genome. The most important representatives of this category were the randomly amplified polymorphic DNAs (RAPDs), in which short PCR primers were used, and the amplified fragment length polymorphisms (AFLPs), in which restriction fragments were selectively amplified after the addition of linkers. The advantage

of these markers is that they do not require *a priori* knowledge of primer sequences in the target organisms. However, results based on RAPDs are notoriously difficult to reproduce, so these markers were soon expelled from genetic analyses as unreliable (Figure 5).

The latest genotyping technology made possible a more comprehensive analysis of polymorphisms related to DNA sequence. This category embraces single nucleotide polymorphisms (SNPs) and data produced by DNA-sequencing methods. SNPs have been successfully used in fine-scale mapping, LD mapping – with important applications to disease association studies – and the inference of past demographic events, such as population expansions and admixture. All the same, DNA-sequencing data provide beyond any doubt the most fine-grained genetic information, because they are free from any ascertainment bias. In the context of population genetics, these data are widely used in phylogenetic analyses and neutrality tests.

The arrival of high-throughput SNP genotyping techniques made possible the extensive study of yet another category of molecular marker, the copy number variants (CNVs). CNVs are genomic structural variations caused by a variable number of copies of a particular segment of DNA, its size typically ranging from 1 kilobase to several megabases (Feuk et al., 2006). Genomic segments with variable copy number may encompass parts of genes, include several known genes or reside entirely outside them. The potential role of copy number variation in susceptibility to complex diseases like schizophrenia, Crohn's disease and psoriasis is supported by a growing amount of published work, with

most of the findings technically validated and/or replicated (reviewed by Wain et al., 2009).

The above retrospect underlines how genetic – and genomic – analysis depends on the available technology. Technological advances point towards a continuous improvement of the assessment of genetic variation, particularly in terms of data quantity and analysis speed.

Main types of molecular markers studied

The following paragraphs contain further information about the categories of molecular markers that were primarily used in this work. Each one of them has unique properties that arise from the different mutation mechanisms involved in their generation.

Alu elements

Alu elements are short interspersed elements (SINEs) of approximately 300 bp and owe their name to the recognition site for the restriction enzyme *Alu*I that some members of the family have (Houck el al., 1979). With approximately 1,100,000 copies (Smit, 1996), Alu elements are the largest family of mobile elements in the human genome (more than 10% of its total mass). The appearance and amplification of Alu elements some 65 MY ago coincided with the origin and dispersion of primates (Deininger and Daniels, 1986). Their distribution throughout the genome is not uniform, as they are primarily found in gene-rich regions (Korenberg and Rykowsky, 1988; Chen et al., 2002).

A typical Alu element is a dimer composed by two monomers – A and B (Figure 6). Monomer B is approximately 30 bp longer than A due to an insertion. Both

monomers have a poly(A) tail at the 3' end. Variation in Alu size is primarily owed to variation in length of the poly(A) tail. Alu elements were ancestrally derived from the 7SL RNA gene, which forms part of the ribosome complex (Ullu and Tschudi, 1984) and are thought to increase in number through a process called retrotransposition, i.e. the reverse transcription of an Alu-derived RNA polymerase III transcript. Since Alu elements have no open reading frame, they are thought to borrow the reverse transcriptase, which is essential for their amplification, from long interspersed elements (LINEs; Mathias et al., 1991).

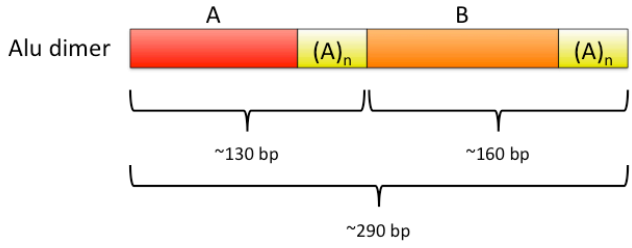


Figure 6: Structural features of a full-length Alu element. The two monomers within the Alu dimer differ in length because of a 32-bp insertion in the B monomer (image adapted from Jobling et al., 2004).

The human Alu family is composed of several subfamilies of different genetic ages (Figure 7). Older Alu subfamilies have the smallest number of family-specific mutations and the largest number of random mutations. Conversely, younger subfamilies have a larger number of family-specific mutations as compared to that of random mutations (reviewed by Batzer and Deininger, 2002). Alu nomenclature is based on family-specific mutations (Batzer et al., 1996; Roy-Engel et al., 2001). There are approximately 5000 human-specific 'young' Alu elements, 25% of which were incorporated into our genome so recently that they are dimorphic for the presence or absence of the insertion.

These dimorphic elements belong to one of several small and closely related 'young' Alu subfamilies, known as Y, Yc1, Yc2, Ya5, Ya5a2, Ya8, Yb8 and Yb9. Individuals can be polymorphic for the presence or absence of these Alu elements at a particular chromosomal location.

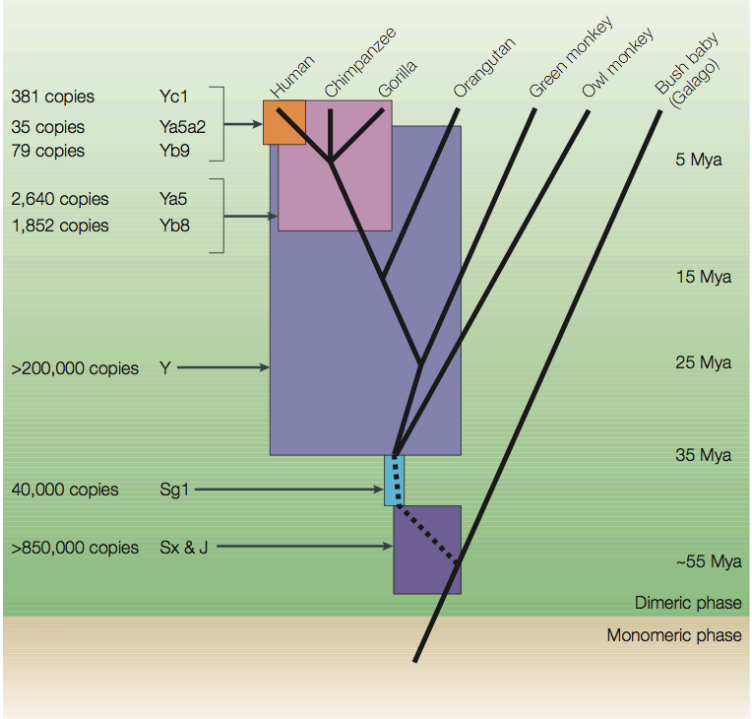


Figure 7: The expansion of the various Alu subfamilies in primates. Colour codes denote the times of peak amplification. Copy numbers of each Alu subfamily are also noted (figure taken from Batzer and Deininger, 2002).

Alu insertion alleles at a given locus are identical by descent, since the probability of two independent Alu insertions occurring at the same genomic position in the human population, given the current low rate of Alu retrotransposition and the relatively short evolutionary time frame involved, is essentially zero. Moreover, the ancestral state of each Alu insertion polymorphism is known to be the absence of the element, as the comparison with other primate DNA sequences showed. Finally, there is no known mechanism that causes reverse mutations (i.e. removal of the element) in Alu

insertion polymorphisms and even when a rare deletion occurs, it always leaves a molecular trace behind (Edwards and Gibbs, 1992). These three properties make Alu elements invaluable for human population genetics and primate comparative genomics.

Although the overwhelming majority of Alu insertions occur in non-coding regions (intergenic regions and introns) with no apparent negative consequences, some Alu elements integrate into the coding or regulatory regions of genes, resulting in several negative effects. Alu insertions account for around 0.1% of all human genetic disorders, such as haemophilia A and breast cancer (Sukarova et al., 2001; Teugels et al., 2005).

Alu elements affect the genome in several ways. Recombination between Alu elements has contributed to the generation of human genetic diversity, through duplications, deletions and translocations, and is responsible for several human genetic disorders (Deininger and Batzer, 1999). Many Alu sequences affect gene expression through changes in their own methylation status, whereas Alu RNA expression potentially influences translation levels (reviewed by Batzer and Deininger, 2002).

Regarding the variation and distribution of the Alu elements in the X chromosome, little was known before their first comprehensive analysis in four human groups (African-Americans, Asians, Europeans and Egyptians; Callinan et al., 2003). The study identified 264 X chromosome Alu elements from eight young Alu subfamilies, 16 of which were found to be polymorphic with various levels of heterozygosity.

Microsatellites

Microsatellites – also known as short tandem repeats (STRs) or simple sequence repeats (SSRs) – are perfect or near perfect tandem iterations of short sequence motifs, typically mono-, di-, tri- and tetranucleotides, although penta- and hexanucleotides are usually classified as microsatellites as well (Figure 8). With more than 1,000,000 loci, microsatellites account for 3% of the genome (Reviewed by Ellegren, 2004). Microsatellites are characterised by remarkably high variability and a remarkably high mutation rate, which are reflected in their multiple alleles and high heterozygosity. Microsatellite variation consists in the number of motifs in each allele rather than in the primary sequence. The advent of PCR in the mid-1980s promoted the successful application of microsatellites in various fields like genome mapping, paternity testing, forensics and the inference of demographic and selective processes (Schlotterer, 2003). Currently, technical problems such as PCR artefacts (stutter bands) complicate the automated scoring of microsatellite alleles.

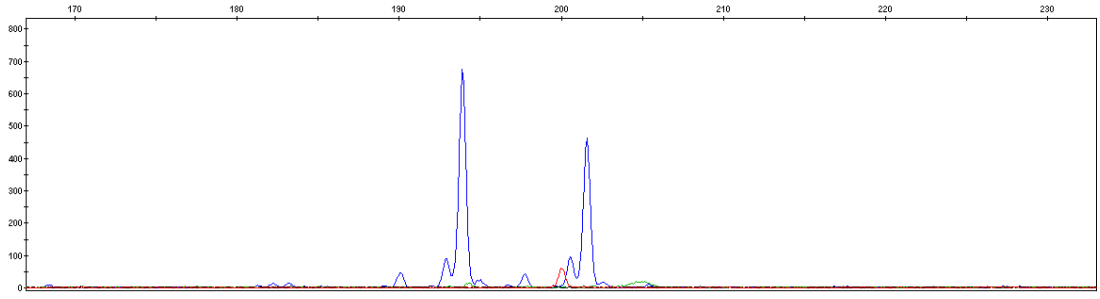


Figure 8: Electropherogram (PCR product size × signal intensity) generated by the GeneMapper® software (Applied Biosystems) displaying the genotype of an individual for microsatellite ss153949702. The two major blue peaks correspond to alleles $(AAAT)_{10}$ and $(AAAT)_{12}$ (for more information see Athanasiadis et al., 2010b).

It is not entirely clear whether microsatellites have a biological function or simply represent 'selfish' junk DNA. Those used as genetic markers in population genetic studies are usually embedded in intragenic or intronic regions and are thus assumed to evolve neutrally, although some exceptions exist: Friedreich's ataxia – an autosomal recessive disease caused by an intronic GAA triplet repeat expansion – is an example of this (Campuzano et al., 1996). In any case, when microsatellites appear in coding regions, there is strong selection against frame shifts, so only trinucleotide repeats are tolerated – provided that they do not substantially alter the function of the protein. Trinucleotides of this kind are involved in human disease. Moreover, recent studies showed that microsatellites are overrepresented in recombination hotspots; there is evidence that certain motifs (Myers et al., 2005) or length differences between sister chromatids in microsatellite loci (Kayser et al., 2006) trigger recombination.

Many theoretical models have been proposed for the evolution of microsatellites. Most of them derive from the 'stepwise mutation model', a symmetric forward-backward random walk of one repeat unit at a time that is independent of repeat length. However, this simple model does not lead to the stationary length distributions observed in real data. As a result, more complex models have been proposed, which impose an upper limit on allele size or introduce a mutational bias such that large alleles mutate preferentially to smaller ones (reviewed by Ellegren, 2004).

In a more recent model, microsatellite allele length is the result of equilibrium between length and point mutations, the former increasing allele length and the latter destroying repeat perfection (Kruglyak et al., 1998). As Ellegren puts it, '*in*

the long run, point mutations break up perfect repeats and reduce the mutation rates of microsatellite loci. Clearly, long microsatellite alleles do not persist indefinitely. However, microsatellite evolution is a dynamic process; therefore, repeats might shrink as well as expand over evolutionary timescales' (Ellegren, 2004). This model explains quite satisfactorily why microsatellites do not expand to enormous numbers.

The mechanism that is thought to produce length mutations in microsatellites is replication slippage – a transient dislocation of the replicating DNA followed by a misalignment (Figure 9). Recombination, albeit crucial to the mutation of minisatellites, does not seem to be involved in microsatellite mutation. As already mentioned, microsatellites are most probably the cause of recombination at certain sites of the genome rather than the consequence of it.

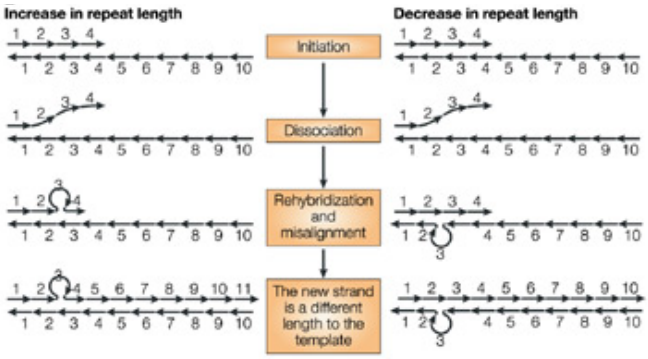


Figure 9: Replication slippage. After the replication of a repeat tract has been initiated, the two strands might dissociate. If the nascent strand then realigns out of register, continued replication will lead to a different length from the template strand. Misalignments introducing a loop on the nascent strand lead to an increase in repeat length. On the other hand, a loop that is formed in the template strand leads to a decrease in repeat length (taken from Ellegren, 2004).

The above view is supported by the observation that in the non-recombining Y chromosome mutation rates are similar to those in the autosomes (Brandström

et al., 2008). In any case, there is no uniform microsatellite mutation rate. Mutation rate is primarily affected by microsatellite length (longer alleles present higher rates), the flanking sequence and possibly the local point-mutation rate, in agreement with the aforementioned Kruglyak model. Sex and age have less effect on microsatellite mutation rate (reviewed by Ellegren, 2004).

Microsatellites – usually $(A)_n$ or other A-rich repeats – are often present close to SINEs and LINEs. As already mentioned, human Alu elements frequently have a structure similar to a mononucleotide microsatellite at the 3' end of their monomers, probably deriving from their insertion mechanism. An alternative theory holds that insertion of a SINE or a LINE might be facilitated by a pre-existing microsatellite at a given site (Wilder and Hollocher, 2001).

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are the simplest type of molecular marker. They consist in singular base substitutions, in which one base is exchanged for another. When the exchange takes place between two purines (i.e. A and G) or two pyrimidines (i.e. C and T), they are called transitions. When a purine is exchanged for a pyrimidine, or vice versa, they are called transversions. Single base insertions or deletions are also included in SNPs, although the mechanisms that give rise to them are different from those of base substitutions. The two fundamental processes of base substitutions are misincorporation of nucleotides during replication and chemical or physical mutagenesis.

Mutation rates of base substitutions are in general so low (roughly ranging between 10^{-8} and 10^{-5}; reviewed by Jobling et al., 2004) that a given mutation at

a given position is unlikely to have recurred or reverted over the time-scale of the evolution of modern humans, so, with some notable exceptions, no independent occurrences are found in appreciable frequencies. A low mutation rate implies that SNPs are identical by descent: two individuals bearing the same allele on a locus have inherited it from a common ancestor. The direction of evolutionary change can usually be established by examining the homologous sequence in the DNA of our closest living relatives, the great apes. Low mutation rate also means that SNPs are primarily biallelic markers.

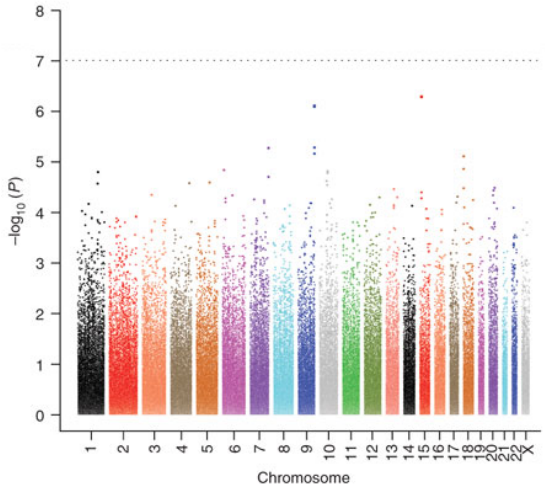


Figure 10: Manhattan plot of the P values in a genome-wide association study of pancreatic cancer. The x axis represents chromosomal locations and the y axis shows P values on a logarithmic scale. Each dot corresponds to a SNP tested for association (taken from Amundadottir et al., 2009).

The ubiquitous presence of SNPs across the genome (one SNP every 100 to 300 bases along the 3-billion-base human genome) and their high potential for automated high-throughput analysis has in recent years turned them into the most commonly used molecular marker for different kinds of genetic analysis, such as disease association and population genetic studies. In genome-wide association studies the human genome is typically scanned for regions associated with complex diseases (Figure 10). Good knowledge of the patterns of SNP variation and linkage disequilibrium is essential. To this end, the ambitious

and hotly debated HapMap project (International HapMap Consortium, 2005) provides a SNP-based haplotype map of the human genome, in which common patterns of human variation are described.

A SNP marker focuses on a specific nucleotide position in the genome, which requires *a priori* knowledge of allelic variation at that position. Whole shotgun genome sequencing of pooled DNA taken from donors provides such knowledge. However, the technique leads to a considerable ascertainment bias in allele frequency and LD patterns, caused by the fact that DNA donors represent only a fraction of human diversity. Fortunately, recent statistical methods successfully address the issue.

HUMAN POPULATIONS IN THE MEDITERRANEAN

Although this work explores the genetic variation of worldwide population samples, emphasis was given in the Mediterranean region. For this reason, this chapter centres around the history and genetic studies related to this region.

Migration in the Mediterranean

The Mediterranean Sea combines a set of characteristics making it one of the most important scenes of human history. Bringing together three continents and having nurtured some of the most brilliant civilizations of antiquity, the Mediterranean has never ceased being one of the most densely inhabited regions in the world. The great demographic events unfolding here have left their mark in palaeontological and archaeological remains, as well as in the genetic composition of the current populations.

Even though an exhaustive account of historical facts is out of the scope of this work, some comments on the history of the Mediterranean may be helpful in understanding this work. Therefore, this section contains a brief outline of the most important migratory events that are thought to have affected the genetic composition of human populations in the Mediterranean.

The prehistoric colonisations

According to the fossil record and archaeological findings, the colonization of Europe by anatomically modern humans occurred around 40,000 years ago. These Upper Palaeolithic people were already present in West Asia by 47,000 years ago and brought their technologies to Europe through the Near East towards the Balkans and from there to the west (Figure 11). Similar dispersals

from the Near East also occurred into North Africa, as confirmed by mtDNA data (Olivieri et al., 2006). At that time, humans were organised in hunting-gathering societies and were already manifesting all the distinctive features of fully 'modern' cultural behaviour including the fabrication and use of sophisticated bone tools and personal ornaments, as well as the creation of remarkably varied and sophisticated forms of both abstract and figurative art, altogether corresponding to the term 'Aurignacian technologies' (reviewed by Mellars, 2004).

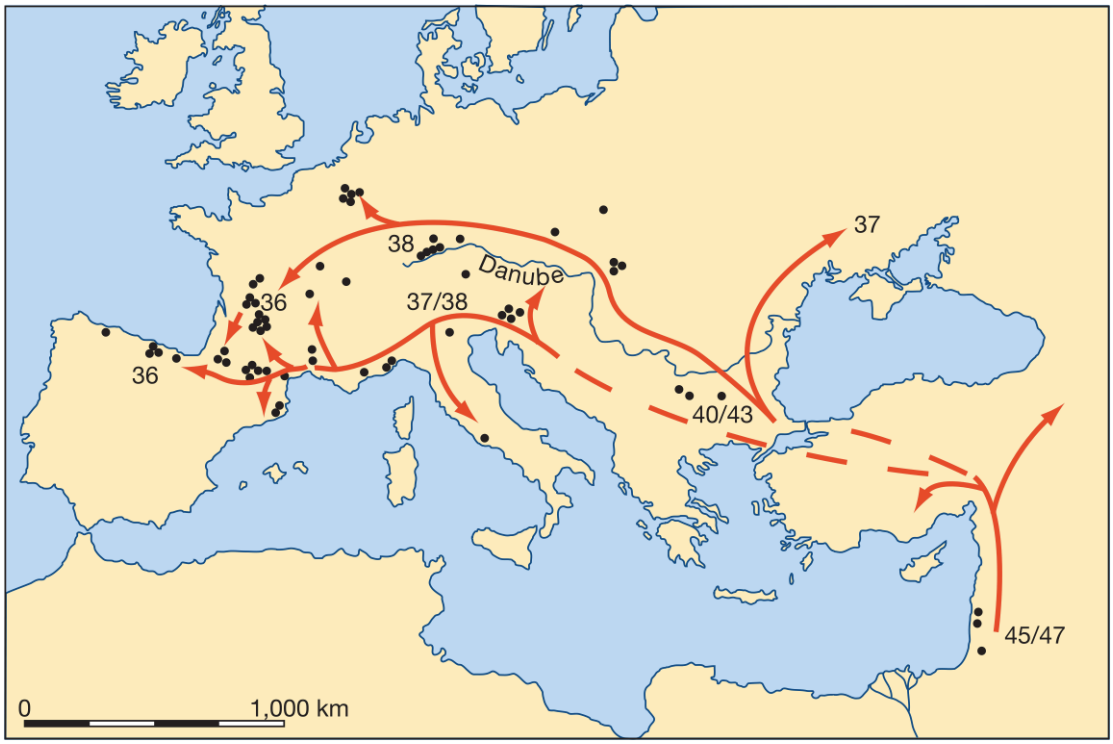


Figure 11: Dispersal routes of the earliest anatomically modern populations across Europe, as reflected in the archaeological data. The southern route represents the colonization of the Mediterranean and corresponds to the 'proto-Aurignacian' technologies. Numbers correspond to dates in thousands of years before present; dashed lines indicate uncertain routes; dots correspond to Aurignacian split-base points (image taken from Mellars, 2004).

The next major change occurred in the wake of the Last Glacial Maximum, around 19,500 years – 25,000 ago. During this time, human populations increasingly sought refuge in Southwest Europe; along the Mediterranean; in the Balkans and the Levant; and on the East European plain. From these refugia human populations re-expanded into central and Northern Europe, as several mtDNA-based lines of evidence suggest; according to recent analyses, these expansions were postglacial (11,000-11,500 years ago) rather than Late Glacial (reviewed by Soares et al., 2010).

The beginning of the Neolithic Age was marked by another technological advance, that of agriculture – undoubtedly one the most important events in human history; with the domestication of plants and animals, humans shifted radically from nomadic, hunting-gathering societies to static, small-scale, family-based communities.

The 'agricultural revolution' began approximately 10,000 years ago in the Near East (Cavalli-Sforza et al., 1994). The diffusion of agricultural technology is currently a hotly debated topic in the field of human evolution. In general, there are two rival models: demic vs. cultural diffusion. According to the demic diffusion model, the agricultural expansion was produced by progressive migrations of agricultural groups from the Near East, following a similar route to Palaeolithic colonization (Figure 11) and substituting the native population as they advanced. Conversely, the cultural diffusion model suggests that diffusion took place through the exchange and assimilation of new technologies between neighbouring groups, without any substantial population movements.

Cavalli-Sforza's synthetic maps based on classical markers revealed a cline of allele frequencies centred in the Near East (Cavalli-Sforza et al., 1994; Figure 12). This finding was interpreted as supporting the demic diffusion model, also supported by similar clinal patterns in the study of some highly variable autosomal loci (Chikhi et al., 1998). However, the results based on other fractions of the human genome (i.e. mitochondrial and Y chromosome DNA) were contradictory, thereby triggering a passionate debate between 'demists' and 'acculturationists' (e.g. Barbujani et al., 1998; Semino et al., 2000; Torroni et al., 2001). In any case, the overall data seem to support an intermediate model of agricultural diffusion, in which population substitution was more intensive in the Southeast and Central Europe, while cultural diffusion played a more important role in Western Europe. It this light, it is worth noting that Cavalli-Sforza's original model assumed a mixed cultural-demic diffusion with emphasis on the genetic contribution of the farmers.

The agricultural expansions of the Neolithic are beyond any doubt the first big migratory event in the most recent history of the contemporary Eurasian populations. As the bearers of the new technology advanced, the genetic diversity of most of the indigenous hunting-gathering populations must have been dramatically affected. However, it is believed that some of these indigenous populations preserved part of their original genetic diversity due to a series of isolating mechanisms. Some well-known Mediterranean isolates are the Sardinians (Angius et al., 2002), the Corsicans (Latini et al., 2004) and populations from the Balearic Islands (Picornell et al., 2005), but maybe the best-studied population in this context are the Basques.

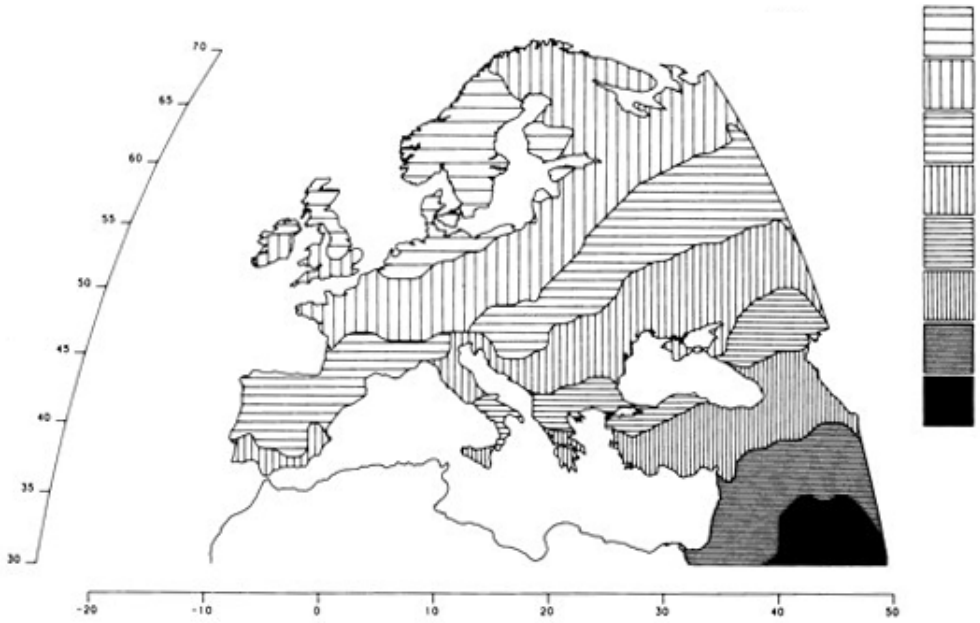


Figure 12: Synthetic map of Europe and Western Asia obtained using the first principal component of classical genetic data (image taken from Cavalli-Sforza et al., 1994).

The Basques are traditionally, but not without controversy, classed as one of the outliers within the European gene pool (Cavalli-Sforza, 1998). Much of their special status is owed to their linguistic particularity: the Basque language is a remarkable non-Indo-European isolate within the relatively homogenous Indo-European landscape (Chen et al., 1995), a fact that many take as a token of 'resistance' to the Neolithic Wave, which granted to the Basques the characterization 'Paleolithic relic'. When it comes to their genetic profile, the Basques present some unusual allele frequencies for some genes. For instance, they have the highest known frequency of the Rhesus negative blood group (Mourant et al., 1976). However, as we shall see further down, a growing number of genetic studies show that the Basques present a degree of genetic differentiation not much more different than that of other European populations.

As for North Africa, the data point towards an agricultural diffusion process parallel to that in South Europe: demographic events must have been more important in Northeast than in Northwest Africa, where the principal factor of expansion must have been cultural diffusion (Bosch et al., 1997; Flores et al., 2000; Esteban et al., 2004).

The great trading colonies of antiquity

With the advent of the Bronze Age in the Near East around 3300 BC, the first trade routes appear in the Mediterranean. The Minoan civilization in Crete played a fundamental role in the spread of the systematic use of bronze to the whole of Western Europe through a far-ranging trade network. Four centuries after the catastrophic end of Knossos, most probably triggered by the eruption of Thera around 1600 BC, the Phoenicians enter the scene. This seafaring people was organised in independent city-states like Tyre, Sidon and Byblos in the Eastern Mediterranean coast. From there, they established a network of trading colonies, which brought them into contact with other civilizations. In the Iberian Peninsula they founded the port of Cadiz in the eleventh century BC and co-existed on friendly terms with the Iberians, a native tribe who spoke a non-Indo-European language. The Phoenicians controlled the trading activities in the entire Mediterranean until the end of the eighth century BC, when they were overshadowed first by the Assyrians and afterwards by the Greeks.

One of the first and most powerful Greek city-states was Korinthos, which rapidly gained control of the trade routes to Italy and founded colonies as far away as Syracuse in Sicily and Apollonia in Libya. By the sixth century BC, the Greeks had colonised the entire Eastern Mediterranean as far west as Sicily. The

only Greek colony in Spain of which there is firm archaeological evidence was Emporion, now Empuries on the Catalan coast.

In about the same period the Western Mediterranean was controlled by Carthage, founded in North Africa in the late ninth century BC by Phoenicians, through its commercial ports in South Spain, the Balearic Islands, Sardinia, Corsica, Sicily and North Africa.

Both Greeks and Carthaginians finally succumbed to the ever-growing power of Rome, which was to establish in the Mediterranean (Mare Nostrum) between 200 BC and 200 AD one of the greatest commercial networks in ancient history. The successor to Rome as the capital of the Roman Empire was Constantinople (Istanbul), founded by Constantine the Great in 330. Constantine's undivided rule did not last long and by 395 the Empire split (for more details see Norwich, 2007).

In North Africa, the progressive drying of the regional terrestrial ecosystem, which started around 6,000 years ago, marked the transition from a 'green Sahara' to the present hyperarid desert – an important geographic barrier in the South of the Mediterranean. It is believed that Sahara's desert ecosystem as we know it today, was established around 2,700 years ago (Kröpelin et al., 2008) and resulted in a more significant isolation, as compared to South Europe, of the native tribes (the Berbers) from the rest of the African continent. However, this isolation never came to be complete, as contacts with sub-Saharan Africa through the coastal routes and oases never ceased to exist. Apart from these sub-Saharan influences, the Berbers were in continuous contact with the Phoenicians and Carthaginians, and later with the Greeks and the Romans.

The 'Barbarian' invasions

Historically, many peoples have dismissed alien cultures and rival civilizations as barbarians. Throughout Roman history this pejorative term was primarily assigned to the Celtic and Germanic tribes that lay in the north of the Roman Empire. The first clear signs of the Celtic culture appeared in the early European Iron Age (eighth to sixth century BC) in Central Europe, with the Hallstatt culture. By around 450 BC, the Celts had expanded over a wide range of lands, going as far west as the Iberian Peninsula, where they shared territory with the native Iberians. The conquest of the Celtic Gaul and Iberia under the Roman Emperor Julius Caesar by the end of the first century BC led to the Romanization of the Celts (for further information see Cunliffe, 1999). The genetic imprint of the Celts in the Iberian Peninsula could be reflected in the high frequency of haplogroup R1b – the most common Y-chromosome haplogroup in Western Europe (Rosser et al., 2000; Scozzari et al., 2001; see Figure 13).

The period between 300 and 700 AD in Europe is known as 'the Barbarian Invasions'. During this period a series of Germanic tribes entered the Roman Empire from northern Europe in consecutive migratory waves. The most important were the Vandals and the Goths. Around the middle of the fourth century, the Vandals fled westwards from the raiding Huns and in 409 they settled in Spain after invading France. There they remained until 428, when the newly crowned King Gaiseric led all his people, some 180,000 people, to North Africa. From there they conquered the Balearic Islands, Corsica, Sardinia and Sicily.

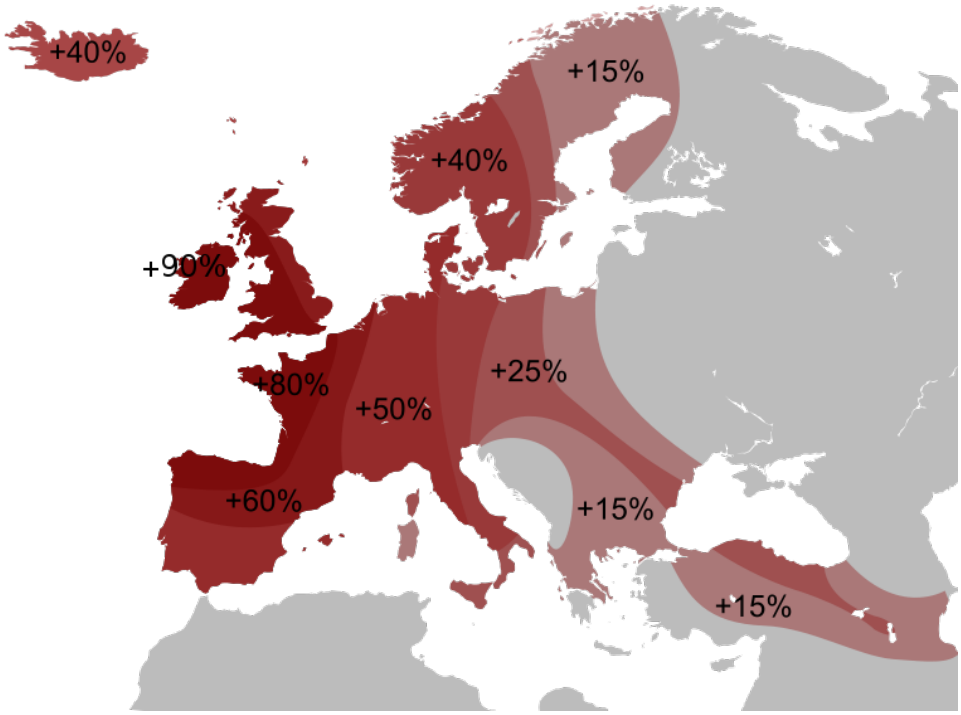


Figure 13: Distribution of the Y chromosome R1b haplotype in the European continent.

The Visigoths, on the other hand, were actually the first Germanic tribe to enter Roman territory in 376 BC, on condition that they would protect the Roman frontier at the Danube. However, they rebelled, invaded Italy in 401 and sacked Rome in 410, before settling in the Iberian Peninsula where they established a relatively stable reign until the invasion of the Arabs (for further information see Norwich, 2007). As for the Ostrogoths, they established their reign in the Italian Peninsula in the late fifth century with the consent of the Roman Emperor Zeno. Italy was later invaded by yet another Germanic tribe, the Lombards, in the sixth century, followed by a West Germanic tribe, the Franks, in the eighth century.

Interestingly, the migration of the Germanic tribes to the Mediterranean does not seem to have had any strong genetic impact at all: following the distribution of a typically 'Germanic' variant of haplogroup R1b, i.e. haplogroup R1b-U106, it can

be verified that, although this haplogroup presents high frequencies in England (21.4%); Germany (20.5%); the Netherlands (37.2%); Denmark (17.7%); and Austria (22.7%), it is practically absent from the Italian (3.5%) and Iberian Peninsula (5%; Myres et al., 2007), suggesting that the Germanic people most probably did not replace the existing populations of the south.

The Arab invasion

The Arab invasion of the North of Africa and the Iberian Peninsula was one of the last great migratory events that are thought to have contributed to the current genetic composition of populations in the Western Mediterranean. There is evidence for the existence of Arab tribes in the desert of Syria and the Arabian Peninsula as early as the ninth century BC. However, the great expansion of this people did not occur until the appearance of Islam, proclaimed by Prophet Mohammed in the early sixth century.

Under the spiritual guidance of Mohammed, the Arabs expanded to Damascus, Jerusalem, Syria, Palestine, Egypt and Persia. By the end of the seventh century they reached the Atlantic, having conquered the Berber Maghreb, and were ready to cross the Strait of Gibraltar into Spain. Within just two years (711-713) they practically controlled the whole Iberian Peninsula, with the exception of Cantabria and the western Pyrenees. During the almost 750 years of Muslim occupation, the Spanish territory was shared among its old inhabitants –Romans and Visigoths – and the newcomers – Arabs and Berbers (for further information see Norwich, 2007).

Modern North African populations are mainly a mixture of native Berbers and Arabs. After converting to Islam, most Berbers adopted the Arabic language and the Muslim culture. Marriages between these two groups were common and have probably contributed to an increased gene flow, which smoothed the genetic differences between the two cultures. However, some significant and relatively isolated Berber groups still exist in North Africa, having conserved their Berber language and culture. In such cases, the genetic differentiation among these particular groups and other North African samples is notable (Fadhlaoui-Zid et al., 2004; Esteban et al., 2006).

Genetic studies in the Mediterranean

History, archaeology and palaeontology have traditionally been the primary tools for the historical reconstruction of the peopling of the Mediterranean. However, as migratory events leave their mark on the genetic composition of the populations, molecular techniques have also been implemented in order to shed more light on several issues regarding the colonization of the Mediterranean. Human population genetics can address some interesting questions regarding prehistoric and historic events, like the relative importance of cultural and demographic diffusion to the expansion of new technologies or the intensity of the contact between different populations.

The following paragraphs make reference to several studies of genetic variation and some highly debated anthropological controversies in the Mediterranean. Emphasis is given to the Western Mediterranean and, more precisely, to those studies that use the same or similar sample collections to those used in this work. These studies can be divided in two categories:

- Genetic studies based on neutral markers, whereby the goal is to investigate the genetic structure and the evolutionary history of human populations, as well as the way these populations are related to each other.
- Genetic studies with an epidemiological interest, whereby the goal is to describe the population-wide distribution of disease-associated polymorphisms, as well as to interrogate their role as determinants of complex diseases.

Genetic studies based on neutral markers

The first human phylogenetic trees were published in the late 1980s and centred on the mtDNA and the Y chromosome (Cann et al., 1987; Lucotte et al., 1989). Because of the relative ease in their implementation, these two fractions of the human genome never ceased being used in human genetic studies ever since, making a detailed citation of all of them too painstaking. As a few examples, we could mention the use of the Y chromosome (Quintana-Murci et al., 2003; Capelli etl a., 2005), as well as the mtDNA (Plaza et al., 2003; Picornell et al., 2005; Torroni et al., 2006; Cruciani et al., 2007) to study the genetic structure in the Mediterranean.

Apart from the uniparentally inherited markers mentioned above, research based on autosomal markers has increased in recent decades. The global distribution of various human blood groups (including the Mediterranean region) was the object of several massive studies in the 1970s and the 1980s (Mourant et al., 1976; Tills et al., 1983). Moreover, as already mentioned, the distribution of more classical markers is reported in Cavalli-Sforza's influential

study in 1994 (see previous chapter). These studies are nowadays considered to be classical references of human population diversity around the globe.

Regarding the Iberian Peninsula, most studies of genetic diversity and population structure in this region include a sample from the Basque Country. Interestingly enough, however, in a study of three Spanish samples based on classical markers the most differentiated sample actually came from Pas Valley rather than from the Basque Country (Esteban et al., 1998). A Y-chromosome microsatellite study revealed a southwest-northeast gene flow within the Iberian Peninsula, as well as a significant level of heterogeneity in the Basques (Peña et al., 2006), with the population from Northern Navarre presenting the most differentiated allele frequencies of classical markers as compared with other Basque groups (Calderón et al., 2006).

Controversy is not uncommon in the case of the Basques: when a set of 13 microsatellites was analysed in two populations from the Basque Country, the samples presented certain level of stratification and were clearly separated from other European and North African populations (Pérez-Miranda et al., 2005). However, when 8 autosomal Alu polymorphisms were analysed in the same two populations, the data showed no substantial genetic heterogeneity either between the two samples or between the Basques and any other Europeans (García-Obregón et al., 2007).

As for population structure in Northwest Africa, there seems to be no general agreement here either: a microsatellite study showed that the genetic differentiation among Northwest African populations was very low, leading to the conclusion that the arabisation of the region was mainly a cultural process

(Bosch et al., 2000). Similarly, the analysis of 5 autosomal Alu polymoprhisms in one Arab and one Berber group from the Tunisian island of Djerba revealed a homogeneous distribution of Alu insertions in both groups, reflecting ancient relationships between them (Ennafaa et al., 2006). However, in a recent X-chromosome study a Moroccan sample was found to be different from all other North African samples considered (Tomas et al., 2008). In any case, an older study based on an extensive collection of data from classical markers revealed a differentiation pattern between Berbers and Arabs from Northwest Africa on the one hand and populations from Libya and Egypt on the other (Bosch et al., 1997), which reflects, despite some contradictions, a more generalised pattern of east-west differentiation in North Africa.

Controversy is even more pronounced when population relationships between the two sides of the Western Mediterranean are investigated. In this context, some scholars see the Strait of Gibraltar as a genetic barrier between North African and South European populations, while others see it as a bridge of cultural and genetic diffusion.

A study based on red cell enzymes detected genetic affinities between an Andalusian sample and some North African populations – a potential consequence of Muslim rule in the South of the Iberian Peninsula (Kandil et al., 1999). In another study based on 8 autosomal Alu polymorphisms and APOE polymorphisms, a low genetic differentiation between North Africans and populations from the Iberian Peninsula was found, again reflecting potential cultural affinities (Bahri et al., 2008). Similarly, an autosomal Alu study showed that the genetic differentiation found between the two sides of the Western

Mediterranean may well be explained by isolation by distance alone (González-Pérez et al., 2003). The 'melting pot' model finds allies in many other studies (Bahri et al., 2008; Izaabel et al., 1998; Arnaiz-Villena et al., 2002; Plaza et al., 2003). Besides, a recent X chromosome SNP study showed that, with the notable exception of the Moroccans, the Mediterranean region exhibits a high overall genetic homogeneity (Tomas et al., 2008), which seems to tie in with an apparently weak genetic structure between South Europeans and North Africans, as revealed by an analysis of Y chromosome microsatellites (Quintana-Murci et al., 2003).

However, the genetic differentiation found in other studies points towards the Strait of Gibraltar acting as a barrier to gene flow (Comas et al., 2000, García-Obregón et al., 2006; Bosch et al., 1997; Bosch et al., 2001; Harich et al., 2002a). According to this hypothesis, the high evaporation of the Mediterranean Sea results in the draw of water from the Atlantic Ocean through the Strait of Gibraltar. This fact produces a strong maritime current that might have made navigation difficult and consequently restricted gene flow. Alternatively, the Neolithic advance may have run in parallel along the two Mediterranean shores (Bosch et al., 1997; Simoni et al., 1999), causing a cultural difference by bringing Indoeuropean languages to the Northern Mediterranean shore and Afroasiatic languages to the Southern shore (Renfrew 1991; Barbujani et al., 1994) capable of keeping gene flow between the two shores low.

Genetic studies with an epidemiological interest

Sometimes population relationships are interrogated using functional (i.e. non-neutral) genetic data of potential interest in epidemiological studies. Bearing in

mind that there are virtually innumerable studies focusing on functional variation in the Mediterranean, this paragraph focuses on those closely related to this work. For instance, genetic relationships among populations from the Iberian Peninsula, North Africa and Sardinia were investigated through the molecular variation in a series of functional genes involved in cardiovascular diseases, revealing a high sub-Saharan gene flow into Morocco (Moral et al., 2003). Another study in a similar set of populations was carried out, this time in the endothelial nitric oxide synthase gene (eNOS), also associated with cardiovascular diseases (Via et al., 2003a). Both studies pointed out the special genetic position of Sardinia. In agreement with Moral et al., 2003, a study of allele variation in the LPA, APOE, APOC1, and APOC2 genes (involved in cardiovascular disease) also suggested a certain degree of sub-Saharan influence on the current Moroccan population (Harich et al., 2002).

The variation of two microsatellites located in exon 1 of the Androgen Receptor (AR) gene (involved in prostate cancer) was analysed in a wide set of Mediterranean populations (Esteban et al., 2006). As above, the two polymorphisms detected a high affinity between the Berbers from High Atlas and the sample from sub-Saharan Africa, suggesting gene flow. What is more, from the Iberian samples, the South Spaniards present the highest affinity with the North African samples, in full agreement with Kandil et al., 1999. The AR gene was involved in yet another study of the same populations, in which the distribution of a SNP located between the two previously mentioned microsatellites was found to present high diversity (Esteban et al., 2005).

As for the field of genetic epidemiology, relevant research was primarily developed around cardiovascular diseases. For instance, an unmatched case-control study on Tunisian patients with ischemic heart disease was carried out in order to interrogate the potential role of 4 polymorphisms from the apolipoprotein genes in the development of the disease. However, none of the 4 polymorphisms showed any significant differences between patients and controls (Bahri et al., 2008).

Another epidemiological study involved polymorphism E65 K from the KCNMB1 gene, which codes for a substantial subunit of the Ca^{2+}-dependent potassium channel and also affects susceptibility to ischemic heart disease. The study was conducted in 101 family trios from Spain through a transmission disequilibrium test (TDT, see next section) and detected no association with the disease (Via et al., 2005). In another TDT study on the same 101 families, several eNOS (endothelial nitric oxide synthase) mutations were tested for association with the same disease, again with negative results (Via et al., 2003b).

POPULATION GENETIC STUDIES AND CARDIOVASCULAR DISEASE: EPIDEMIOLOGICAL APPLICATIONS

Studies of the genetic basis of complex diseases: general concepts

Most common diseases are complex, determined by environmental and genetic factors. Complexity arises not only from the total number of factors affecting the manifestation of the disease, but also from the interactions among them (Figure 14). Genetic studies of complex diseases roughly fall into two categories, linkage studies and association studies. The following paragraphs make some reference to the basic concepts behind each category.

Linkage studies

In linkage studies, the evidence for the implication of a locus in a disease comes from the co-inheritance of genetic markers and phenotypes in pedigrees over several generations. There are two major groups of linkage studies: parametric and non-parametric. Parametric linkage studies are typically implemented in the study of Mendelian characters and they require a precise genetic model of inheritance, allele frequencies and penetrance of each genotype. Non-Mendelian characters, on the other hand, are studied through non-parametric designs. Once linkage has been found, additional genetic markers are selected throughout the candidate region to fine-map the trait. Because linkage studies are axiomatically family-based, they are advantageously not susceptible to population stratification – unlike almost all kinds of association study. However, linkage studies have received strong criticism, as their results are notoriously difficult to replicate for low-penetrance mutations.

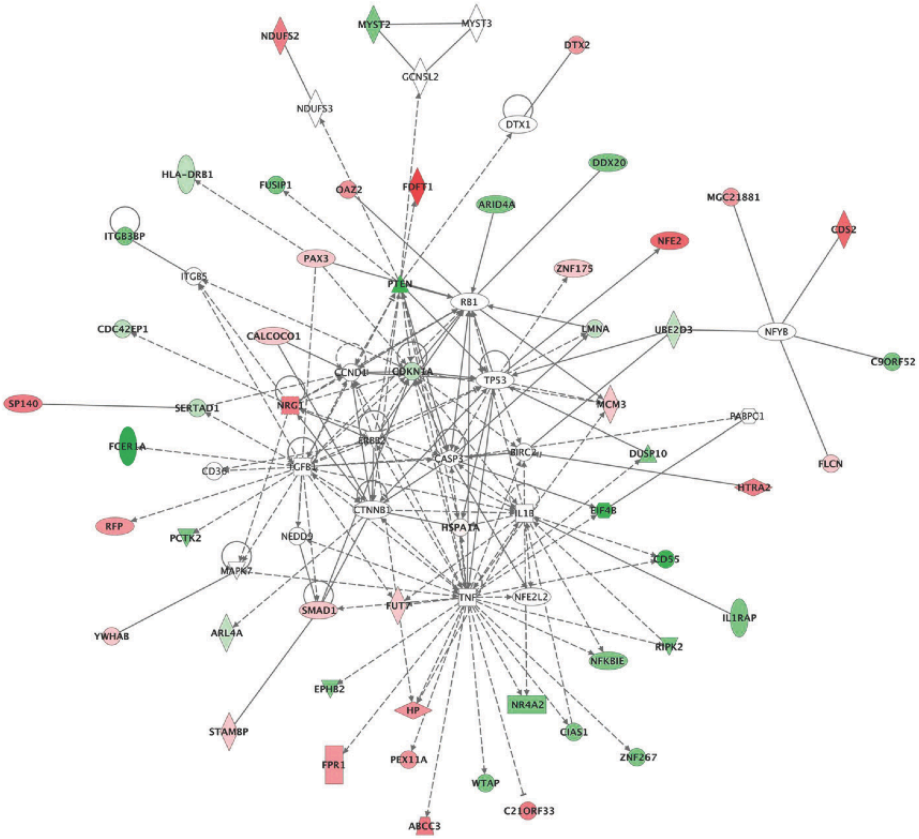


Figure 14: Elements of a genetic network of 200 diabetes-related genes correlating gene expression and human variation; red and green indicate, respectively, genes with positively or negatively correlated expression (image taken from Weiss, 2008; courtesy of Joanne Curran, Southwest Foundation for Biomedical Research, unpublished research).

Association studies

Association studies, on the other hand, essentially look for correlations between phenotype and genotype. The aim is to demonstrate that a particular allele or genetic marker (typically a SNP) is associated with a pathological phenotype. A disease can be associated either with functional genetic variants that have a deleterious effect or, most commonly, with neutral variants that are in linkage disequilibrium with these functional variants. This is typically done through the comparison of a case (i.e. patients) with a control (i.e. healthy individuals) population, although family-based association studies are quite common as well.

Traditionally, association studies have been hypothesis-driven; a candidate gene with a potential implication in the pathophysiology of a disease is tested for association in case-control or family samples. Thus, it can be said that candidate-gene studies build on existing knowledge. However, association studies alone do not provide any causal mechanisms between the genetic marker and the disease, thereby the relevance of the association remains uncertain in the absence of clinical data.

Family trios: a special case of association studies

As already mentioned, family-based association studies stand as an interesting alternative to case-control designs, as they are stratification-free. A common approach is through family trios, i.e. families consisting of the two parents, regardless of their status, and one affected child. The association of a genetic marker with a trait in family trios is typically studied with a transmission disequilibrium test (TDT; Spielman et al., 1993).

The TDT evaluates whether the proportion of transmitted alleles from heterozygous parents to the affected offspring deviates from the expected 50% under the assumption of Mendelian inheritance.

If in a biallelic locus A,a allele *a* is suspected of association with a disease, in heterozygous parents, there are two possibilities of transmission: suspected allele *a* is transmitted while unsuspected allele *A* is not; or unsuspected allele *A* is transmitted while suspected allele *a* is not (Figure 15).

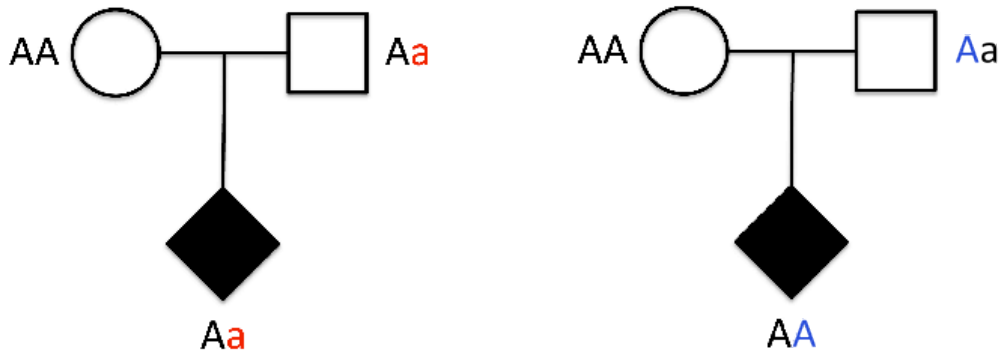


Figure 15: Two informative family trios consisting of an affected child and his/her parents. On the left, the father transmitted one suspected allele to his child without transmitting the unsuspected allele. On the right, the opposite situation is observed.

TDT essentially calculates the frequency in which heterozygous parents transmit the suspected allele to their affected offspring ($n_{a,A}$), with respect to the frequency in which they transmit the other allele ($n_{A,a}$). It follows that, if one allele is transmitted, the other one is not. This is resumed in the following table:

		Non-transmitted alleles	
		A	a
Transmitted alleles	A	$n_{A,A}$	$\mathbf{n_{A,a}}$
	a	$\mathbf{n_{a,A}}$	$n_{a,a}$

The hypotheses tested in TDT are:

$$H_0 : \frac{n_{A,a}}{n_{A,a} + n_{a,A}} = \frac{1}{2} \quad \text{vs.} \quad H_1 : \frac{n_{A,a}}{n_{A,a} + n_{a,A}} \neq \frac{1}{2}$$

The standard McNemar test statistic (McNemar, 1947) for this problem is given by the following formula:

$$T_{TDT} = \frac{\left(n_{A,a} - n_{a,A}\right)^2}{n_{A,a} + n_{a,A}}$$

Under the null hypothesis of equal transmission from heterozygous parents, the T_{TDT} statistic is asymptotically χ^2 distributed.

If a spurious allele has no effect on the disease phenotype, no preferential transmission of it to the affected children is expected. In such cases, transmission of the suspected allele will be observed in 50% of the heterozygous parents. If the analysed allele is significantly transmitted more frequently than expected under Mendelian inheritance, this allele – or another one tightly linked – could indicate susceptibility to a disease.

The pathophysiology of cardiovascular diseases

Cardiovascular diseases (CVDs) are diseases of the circulatory system, i.e. the heart and blood vessels. Although the term technically embraces all cardiovascular disorders, it is mainly used to refer to those that have an atherosclerotic origin.

Atherosclerosis is the condition in which the interior surface of an artery thickens by the formation of atheromatous plaques, leading to blood vessel

constriction. Atheromatous plaques are formed for the most part by the accumulation of cells and the deposition of lipids and calcium. In the process, the arteries lose their elasticity and blood flow is obstructed. The culminating event of atherosclerosis is thrombosis – the formation of clots – by which blood flow is disrupted with severe consequences.

Thrombosis is closely related to haemostasis, a complex process that stops bleeding through the synchronised action of vascular spasms, platelet plug formation (primary haemostasis) and blood coagulation (secondary haemostasis).

In coagulation, clots are formed by the conversion of fibrinogen to fibrin, which strengthens the platelet plug in response to a sequence of enzyme activations know as the 'coagulation cascade'. These enzymes are plasma proteins, called coagulation factors. Coagulation factors are generally indicated by Roman numerals with a lowercase 'a' indicating the active form of the enzyme. Most of these enzymes are serine proteases, which act by cleaving other proteins at specific sites.

The coagulation cascade takes place through two pathways, the contact activation (intrinsic) pathway and the tissue factor (extrinsic) pathway, both leading to fibrin formation. The pathways are series of reactions, in which an inactive enzyme precursor of a serine protease and its glycoprotein co-factor are activated and then catalyse the next reaction in the cascade, ultimately resulting in the formation of fibrin (Figure 16). The risk of thrombosis is determined by abnormalities and quantitative physiological variations in these pathways.

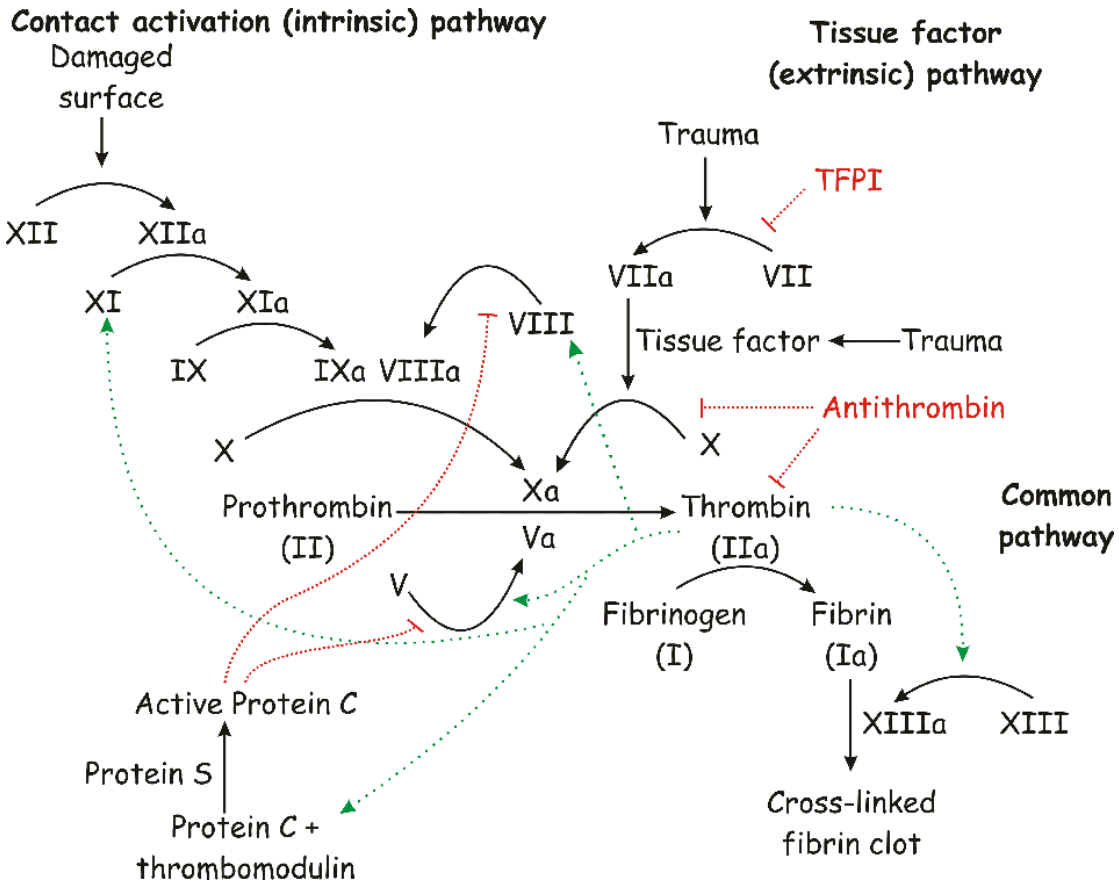


Figure 16: The coagulation cascade.

Ischemic heart disease

Definition

Ischemic heart disease (IHD) is one of the most common cardiovascular diseases. It has an atherosclerotic origin and affects the heart. This alteration is caused by a decrease in blood supply to the myocardium, usually as a consequence of atherosclerosis of the coronary arteries (coronary artery disease). The two most common clinical manifestations of IHD are angina pectoris and the much more detrimental myocardial infarction (heart attack).

Angina pectoris

Angina pectoris is severe chest pain caused by the lack of blood supply to the heart muscle, usually due to the temporary obstruction of the coronary arteries. Most patients with angina complain about chest discomfort rather than actual pain: it is usually described as a pressure, heaviness or tightness, as well as a squeezing, burning or choking sensation. Apart from chest discomfort, pain may also be experienced in the epigastrium, back, neck, jaw or shoulders. Angina is typically precipitated by exertion or emotional stress. Pain may be accompanied by breathlessness, sweating or nausea. It usually lasts for about 3 to 5 minutes, and is relieved by rest or specific anti-anginal medication.

Myocardial infarction

A myocardial infarction (MI) occurs when the interrupted blood supply to the heart causes necrosis of myocardial tissue. Classical symptoms of myocardial infarction include sudden chest pain (typically radiating to the left arm or left side of the neck), shortness of breath, nausea, vomiting, palpitation, sweating, and anxiety (often described as a sense of impending doom). Women may experience fewer typical symptoms than men, most commonly shortness of breath, weakness, a feeling of indigestion, and fatigue.

In the first days after a heart attack, the rate of patient mortality ranges from 25-50%, due to complications that may include congestive heart failure, myocardial rupture, life-threatening arrhythmia, pericarditis or cardiogenic shock. In the following 6 months mortality rate drops to 10% and after a year it stabilises at 3-4%.

Risk factors

Cardiovascular diseases are complex, in the sense that their manifestation is the result of multiple genes in combination with lifestyle and environmental factors. It follows that many factors – genetic and environmental – are expected to affect CVD susceptibility. So far, more than 300 risk factors have been associated with ischemic heart disease and stroke. Most of the relevant knowledge comes from the observations of independent longitudinal studies, like the Framingham Heart Study, which started in 1948 (Kannel et al., 1988), the Seven Countries Study, which dates back to the 1950s (Keys et al., 1986), and the more recent MONICA Project, established in the early 1980s (Tunstall-Pedoe, 2003).

In the Atlas of Heart Disease and Stroke (Mackay and Mensah, 2004), risk factors are classified as modifiable (preventable) and non-modifiable (non-preventable). Interestingly, approximately 75% of CVD can be attributed to conventional risk factors. Some of the most important modifiable risk factors are high blood pressure, abnormal blood lipids, smoking, drinking alcohol, medication for contraception and hormone replacement, physical inactivity, obesity, unhealthy diet, diabetes mellitus, even low socioeconomic status, depression, psychosocial stress. Among the non-modifiable risk factors are advancing age, gender, family history or heredity affecting complex processes like blood coagulation, inflammation, lipid metabolism etc.

Mortality rates

Cardiovascular diseases are the most common cause of death: more people die annually from CVDs than from any other cause. In 2005, CVDs killed ~17.5

million people worldwide, representing 30% of all deaths. Of these deaths, an estimated 7.6 million were due to ischemic heart disease and 5.7 million were due to stroke or other forms of cerebrovascular disease. Over 80% of CVD deaths take place in low- and middle-income countries and occur almost equally in men and women. Statistics are not very optimistic for the medium-term future, as it is expected that by 2020 ischemic heart disease and stroke will become the leading cause of both death and disability worldwide (Mackay and Mensah, 2004). In Europe, CVDs are reported to kill 4.3 million people every year (Allender et al., 2008), a number that accounts for almost half of all deaths in the region (48%).

Haemostasis and thrombotic risk

Susceptibility to thrombosis is typically studied through quantitative traits, usually plasma levels of various haemostatic factors. Although arterial thrombosis is the primary concern in this work, a genome-wide scan demonstrated a substantial overlap in the genetic contribution to venous and arterial thrombosis. This overlap suggests that most genes studied have a pleiotropic effect on thrombotic risk (Souto et al., 2000).

Most of our knowledge of the genetic factors involved in common thrombotic diseases derives from case-control association studies. This design became popular after the identification of the factor V Leiden (FVL) mutation, which is considered to be the paradigm of CVD candidate-gene association studies. Numerous haemostatic factors have been implicated as possible contributors to the thrombotic phenotype. For instance, there is evidence for significant genetic correlations between thrombosis and plasma levels of von Willebrand factor, homocysteine, tissue plasminogen activator, as well as coagulation factors VII,

VIII, IX, XI and XII. All haemostasis-related phenotypes are under substantial genetic control as several family-based, twin and pedigree studies revealed (reviewed by Soria and Fontcuberta, 2005). From the above haemostatic factors, we distinguish coagulation factor VII and XII, as they are the two functional regions studied in this work.

Coagulation factor VII

Coagulation factor VII is a vitamin K-dependent glycoprotein that is synthesised in the liver and secreted into the blood as an inactive zymogen (Fair, 1983). After endothelial damage, tissue factor (TF) is exposed and activates FVII (FVIIa). It is the FVIIa-TF complex that initiates the tissue factor pathway of the coagulation cascade (Figure 16). Many studies have proposed that high FVII plasma levels are related with an increased risk of cardiovascular disease.

Functional FVII plasma levels were shown to present 53% heritability (Souto et al., 2000) and are determined by a single locus, that of the F7 gene (Soria et al., 2005). F7 spans 12.8 kb on chromosome 13q34-ter and contains nine exons and eight introns, coding for a 406 amino acid mature protein.

It has been proposed that variability in FVII plasma levels is the result of regulatory non-coding and intronic variants rather than the result of amino acid changes (Sabater-Lleal et al., 2006).

Coagulation factor XII

Coagulation factor XII is a serine protease precursor, produced and secreted by hepatocytes. FXII is primarily involved in two opposite biochemical processes,

i.e. initiation of the contact activation pathway of the coagulation cascade and fibrinolysis (Kluft et al., 1987).

Low FXII plasma levels have been reported to affect the susceptibility to thrombotic disease as the result of a deficient fibrinolytic system (Foncea et al., 2001; Bach et al., 2008; Doggen et al., 2006). FXII plasma levels present 67% heritability (Soria et al., 2002) and are affected by a region on chromosome 5 where the structural F12 gene has been mapped. The F12 gene consists of 13 introns and 14 exons that cover 12 kb.

Interestingly, a sole SNP in F12 (FVII46C>T) seems to be a major determinant of factor XII plasma levels, accounting for 40% of the variance in FXII activity levels in a Spanish Mediterranean population (Soria et al., 2002). Polymorphism 46C>T is a cytosine to thymidine transition at position 46 from the transcription initiation point in exon 1, in the 5´-untranslated region of the F12 gene. It affects the translation efficiency, probably by creating a second ATG translation initiation codon in a different frame, leading to reduced FXII plasma levels (Kanaji et al., 1998; Ishii et al., 2000).

Since its discovery, many genetic studies have been conducted to interrogate the genetic importance of 46C>T. Studies on Spanish samples showed that genotype T/T is an independent risk factor for venous thrombosis (Tirado et al., 2004), ischemic stroke and coronary artery disease (Santamaría et al., 2004a;b). However, controversial results in a subsequent study (Bach et al., 2008) led Kanaji and colleagues to propose that low FXII plasma levels are not the cause of thrombosis, but rather the result of it (Kanaji et al., 2008).

POPULATIONS STUDIED IN THIS WORK

As already mentioned, this work is primarily focused on human populations from the Mediterranean region with a particular emphasis on the Iberian Peninsula and Northwest Africa (Figure 17). In addition to these populations, three non-Mediterranean populations were also considered in order to contextualise our findings: a sample from the Ivory Coast, as a representative of the sub-Saharan genetic variation, and two indigenous samples from Bolivia, as representatives of the Native American variation. A brief description of these populations is presented in the following paragraphs.

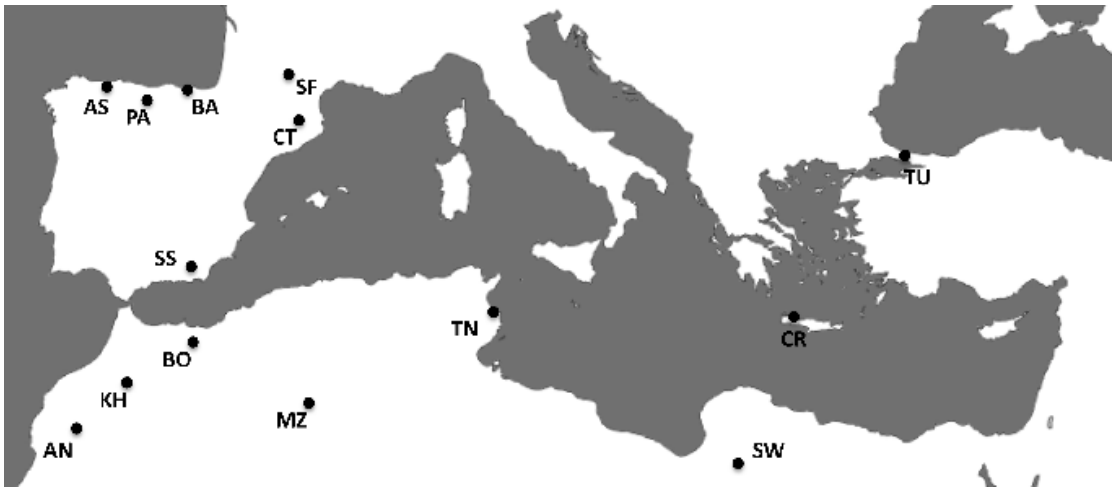


Figure 17: Geographic location of the populations studied around the Mediterranean. AN: Asni, High Atlas, Morocco; AS: Oviedo, Asturias, Spain; BA: Basque Country, Spain; BO: Bouhria, Northeast Atlas, Morocco; CR: Crete, Greece; CT: Catalonia, Spain; KH: Khenifra, High Atlas, Morocco; MZ: M'zab, Algeria; PA: Pas Valley, Cantabria, Spain; SF: Toulouse, South France; SS: South Spain, Andalusia; SW: Siwa Oasis, Egypt; TN: Monastir, Tunisia; TU: Istanbul, Turkey.

General populations

Southwest Europe

The Iberian Peninsula is of particular interest because of its complex history over the last two millennia, involving the long-term coexistence of several populations with distinct geographical origins and cultural particularities (i.e. Christians, Muslims and Jews). In this work, the Iberian Peninsula is represented by five samples:

- A sample from Asturias (North Spain). Asturias has been occupied by humans since the Lower Paleolithic and was characterised by cave paintings in the eastern part of the area during the Upper Paleolithic.
- A sample from the rural areas of the province of Guipúzcoa (the Basque Country) in the North of the Iberian Peninsula. As already mentioned, the Basques speak a non-Indo-European language, which is believed by some scholars to have contributed to their genetic isolation. In this work we also evaluate this hypothesis.
- A sample from Pas Valley in Cantabria (North Spain). Some scholars consider the 'Pasiegos' to be another example of genetic outlier in the Iberian Peninsula, caused by small population size in the previous centuries and a high degree of endogamy and consanguinity (Sanchez-Velasco et al., 2003).
- A sample from Olot, the capital of the Catalan province of Garrotxa in Northeast Spain, noted for some 40 well-preserved volcanic cones in the western part of the province.

- A sample from La Alpujarra a mountainous region in the provinces of Granada and Almeria (South Spain). This region is thought to have served as a refuge from several invasions, particularly related to the Muslim conquest of the Iberian Peninsula.

France is represented by a sample from Toulouse, capital of the French region of the Midi-Pyrénées in the south of the country.

Southeast Europe

In this work, Southeast Europe is represented by two samples:

- A sample from Crete (South Greece), a mountainous island with a typically Mediterranean climate and a rich history (see previous sections).
- A sample from the city of Istanbul in Turkey, which consisted of students from the Fatih University of Istanbul with at least three generations of proved ancestry in this geographical region.

North Africa

As already mentioned, the overall North African population is primarily comprised of two populations with different geographic origins and cultural backgrounds, the Arabs and the Berbers. Whereas the Berbers are considered to be the autochthonous inhabitants of North Africa, the Arabs arrived many centuries later (see previous section). In this work, North Africa is for the most part represented by Berber samples:

- A Berber sample from Asni in High Atlas (Morocco). This region has been relatively isolated owing to the difficult access and the harsh climate. Asni

is actually a group of villages scattered over the area at altitudes exceeding 1,900 m.

- A Berber sample from Khenifra in Middle Atlas (Morocco). Khenifra is considered to be the capital of the Berber tribes in the Middle Atlas. The sampled individuals come from four main tribal groups: Zayane, Ichkern, Ait Sgougou, and Beni M'Guid.
- A Berber sample from Sidi Bouhria, a town located in the Berkane province in Northeast Atlas (Morocco). The Berbers of this sample form part of a large tribe – the Beni Snassen.
- A sample of M'zab Berbers from Algeria. These Berbers present some linguistic particularities in their spoken and written language, as well as in their social and cultural identity as compared to the other Berbers, possibly leading to their isolation evidenced in several genetic studies (Bosch et al., 2000; González-Pérez et al., 2003).
- A Berber sample from Siwa Oasis in Egypt. This population represents a good example of geographic isolation among the Berbers, with a high degree of endogamy and certain genetic particularities on a level of both classical (Amory et al., 2004) and DNA markers (Esteban et al., 2005).

The only non-Berber North African sample used in this work is an Arab-speaking sample from the city of Monastir in Tunisia.

The Ivory Coast

Evidence from genomic variation suggests that anatomically modern humans arose in a single region somewhere in Northeast Africa about 100,000-200,000 years ago (Jobling et al., 2004). As a consequence, all human populations that

inhabited the rest of the world carry a subset of the genetic variation found in sub-Saharan Africa through a succession of founder effects and expansions (Weiss and Long, 2009). It is therefore particularly useful for population genetic studies to include a representative of the sub-Saharan genetic variation. To this end, this work includes a sub-Saharan sample from the Ivory Coast in West Africa. The sample corresponds to the ethnic group of Akans (or Ahizi/Aizi) from the village of Nigui-Saff, located in the surroundings of the capital of the country.

Native American Bolivia

Archaeological and historical records suggest that modern Bolivian populations are the result of complex historic interactions among people of different languages and cultures. Most data point to the Central Andes (i.e. the Bolivian Altiplano and Peru) as the heartland of the first complex societies of South America. In this work we considered two populations, the Quechuas and the Aymaras, corresponding to the two main Native American linguistic groups of the area. The Quechua language is spoken by 12 million people in Ecuador, Peru, Southern Bolivia and Northern Chile, while the Aymara language has 1.5 million speakers mainly in Bolivia.

Samples for epidemiological studies

Family trios from Spain

A total of 101 families (302 individuals) were recruited from four hospitals in Barcelona, Spain. Each family unit consisted of one individual diagnosed for ischemic heart disease (IHD) and both parents. In the cases where only one or neither of the parents was available, siblings were recruited in order to

reconstruct the parental genotype. The criteria of the recruitment of a subject in the study were an age less than 55 and a diagnosis of myocardial infarction or angina pectoris. Myocardial infarction was diagnosed by clinical, enzymatic, and electrocardiography alterations that are typical in a myocardial ischemic injury and necrosis; angina pectoris was diagnosed by a coronariography with an obstruction of more than 50% in a main coronary artery.

Case-control from Tunisia

The case-control design consisted of 76 patients with ischemic heart disease complicated by myocardial infarction, all from Monastir, and 118 healthy and unrelated autochthonous individuals from North and South-Central Tunisia. The criteria used for the diagnosis of the affected individuals participating in this design were similar to those used in the Spanish family trios.

Goals

GOALS OF THE STUDY

This work deals with the genetic characterisation of human populations based on Alu polymorphisms from the X chromosome, as well as SNPs and microsatellites from two autosomal genomic regions related to the risk of cardiovascular disease (Coagulation Factors VII and XII). The variation found in these regions was further used to address a series of interesting issues in the field of human population genetics. More precisely, the present work was designed to accomplish the following specific goals:

- To describe – for the first time – the variation of 13 polymorphic X chromosome Alu elements in 6 human populations from the Mediterranean region and sub-Saharan Africa in order to assess (i) the genetic relationships between South European and North African populations; (ii) the possible sub-Saharan influences over them; and (iii) the usefulness of these markers in human population studies.
- To analyse the molecular variation in and around the F7 genomic region in 16 human groups from the Mediterranean, sub-Saharan Africa and Native American Bolivia through 5 mutations (4 SNPs and one insertion/deletion polymorphism) from the F7 promoter region, as well as 6 SNPs and 4 novel microsatellites from the flanking region. This variation was further used to (i) explore the linkage disequilibrium patterns in the F7 promoter region and (ii) to shed more light on the evolutionary history of the F7 gene across different human populations.
- To analyse the molecular variation in and around the F12 genomic region in 16 human groups from the same geographical regions as mentioned above

through a cluster of 5 SNPs and 3 novel microsatellites. The variation found was used (i) to explore – together with the variation from the F7 gene – the genetic structure of human populations in the Mediterranean and the role of sub-Saharan gene flow in their differentiation and (ii) to assess the role of the FXII 46C>T polymorphism in the susceptibility to ischemic heart disease in two samples from the Western Mediterranean.

Results

SUPERVISOR'S REPORT ON THE QUALITY OF THE PUBLISHED ARTICLES

The doctoral thesis 'Genetic variation of the X chromosome and the genomic regions of Coagulation Factors VII and XII in human populations: Epidemiological and evolutionary considerations' is based on the original results obtained by Georgios Athanasiadis and published in four international peer-reviewed journals.

In all four publications, genetic variation is used in order to address several issues regarding the demographic and biological history of various human groups. The large amount of data obtained (both in terms of populations and markers) and the variety of sophisticated statistical tests that were carried out are beyond any doubt a considerable contribution to the scientific community.

The importance of the research conducted is demonstrated by the quality of the four journals:

- *European Journal of Human Genetics* is the official Journal of the European Society of Human Genetics, publishing high-quality papers in the field of human genetics and genomics. It is indexed in SCI (Science Citation Index) with a current impact factor of 3.925 and classified in the second quartile of the area 'Genetics & Heredity' (ranking: 35/138);

- *BMC Research Notes* BMC Research Notes is an open access journal publishing scientifically sound research across all fields of biology and medicine, enabling authors to publish updates to previous research, software tools and databases,

data sets, small-scale clinical studies, and reports of confirmatory or 'negative' results. As this is a new journal in the BMC family, it is not yet indexed in SCI.

- *Annals of Human Genetics* is published in association with University College, London. It is indexed in the SCI with a current impact factor of 2.195 and classified in the third quartile of the area 'Genetics & Heredity' (ranking: 84/138);

- *BMC Evolutionary Biology* is an open access journal publishing original peer-reviewed research articles in all aspects of molecular and non-molecular evolution of all organisms, as well as phylogenetics and palaeontology, also indexed in SCI. With an impact factor of 4.050, this journal is in the top quartile of the area 'Genetics & Heredity' (position 29/138).

Signed by Dr. Pedro Moral Castrillo

Barcelona, 9 April 2010

Results I

Athanasiadis et al., 2007

The X chromosome Alu insertions as a tool for human population genetics: data from European and African human groups

Georgios Athanasiadis, Esther Esteban, Marc Via, Jean-Michel Dugoujon, Nicholas Moschonas, Hassen Chaabani and Pedro Moral

European Journal of Human Genetics 2007; 15(5): 578-583

Resumen en castellano

Las secuencias Alu son secuencias cortas repetidas dispersas por el genoma (del inglés: short interspersed nuclear elements – SINEs) de una longitud alrededor de unos 300 pares de bases. Con aproximadamente más de un millón de copias, las secuencias Alu son los elementos móviles más abundantes en el genoma humano. Algunas de las Alu son polimórficas en el sentido de que la inserción de la secuencia no está fijada; un cromosoma puede bien portar la inserción, bien no.

Dichos polimorfismos Alu resultan de gran utilidad para los estudios antropológicos que investigan el origen y las relaciones genéticas entre diversas poblaciones humanas, debido a dos importantes ventajas sobre otros marcadores genéticos: por un lado, los portadores de una inserción en un locus determinado son idénticos por descendencia y, por otro, el estado ancestral de un polimorfismo Alu puede ser determinado y casi siempre corresponde a la ausencia de inserción.

La variación de las secuencias Alu del cromosoma X en poblaciones humanas se ha descrito anteriormente, pero dichos polimorfismos no se habían usado en estudios poblacionales, representando este trabajo uno de los primeros estudios sistemáticos sobre la variación de las inserciones Alu polimórficas del cromosoma X.

En este artículo se presenta la distribución de frecuencias encontradas en el análisis de 13 polimorfismos Alu del cromosoma X (Ya5DP62, Ya5DP57, Yb8DP49, Ya5a2DP1, Yb8DP2, Ya5DP3, Ya5NBC37, Yd3JX437, Ya5DP77, Ya5NBC491, Yb8NBC578, Ya5DP4 y Ya5DP13) en 6 poblaciones humanas de España (País Vasco), Grecia (isla de Creta), Norte de África (Alto Atlas de Marruecos; Oasis Siwa de Egipto; y Monastir de Túnez) y África subsahariana (Costa de Marfil).

De los 13 polimorfismos Alu estudiados, ocho han mostrado una diversidad génica notablemente alta en todas las poblaciones (heterocigosidad media: 0.209 en Europa; 0.250 en Norte de África; y 0.342 en la Costa de Marfil).

Las relaciones genéticas encontradas concuerdan con un patrón de diferenciación geográfico, con la excepción de la muestra de Túnez, la cual presenta algunas características peculiares: pese a su situación geográfica, dicha población queda lejos de las otras dos poblaciones norteafricanas en el gráfico del análisis de componentes principales. Por otro lado, la isla de Creta y el País Vasco muestran la menor distancia genética (0.016, $p>0.05$), poniendo en evidencia la escasa estructuración poblacional en el sur de Europa.

El análisis con el programa STRUCTURE ha detectado una sutil estructuración poblacional en las 6 muestras, con los norteafricanos presentando una influencia subsahariana más elevada que los europeos.

Los resultados de este trabajo muestran que los polimorfismos Alu del cromosoma X constituyen unos marcadores genéticos fiables y útiles para investigar las relaciones poblacionales a nivel sub-continental.

Supervisor's report on the involvement of the PhD student in the development of this paper

UNIVERSITAT DE BARCELONA

Dr. **Pedro Moral Castrillo**, Professor at the Department of Animal Biology of the University of Barcelona and supervisor of the doctoral thesis "Genetic variation of the X chromosome and the genomic regions of Coagulation Factors VII and XII in human populations: Epidemiological and evolutionary considerations" by **Georgios Athanasiadis**, hereby certifies that the participation of the above student in the article "**The X chromosome Alu insertions as a tool for human population genetics: data from European and African human groups**", published in the *European Journal of Human Genetics,* consisted of the following tasks:

- DNA extraction from the Basque Country sample

- Participation in the design of the study and selection of the analysed markers together with Dr. Pedro Moral Castrillo

- Genotype determination of the Alu polymorphisms in the lab

- Creation of the genotype database and statistical analysis of the data

- Participation in the manuscript drafting together with Dr. Pedro Moral Castrillo

In addition, it is important to note that none of the co-authors of this article have used the results of this work in any implicit or explicit way to develop another doctoral thesis. As a consequence, this article forms part of the doctoral thesis of Georgios Athanasiadis exclusively.

Signed by Dr. Pedro Moral Castrillo

Barcelona, 9 April 2010

European Journal of Human Genetics (2007) 15, 578–583

www.nature.com/ejhg

ARTICLE

The X chromosome Alu insertions as a tool for human population genetics: data from European and African human groups

Georgios Athanasiadis[1], Esther Esteban[1], Marc Via[1], Jean-Michel Dugoujon[2], Nicholas Moschonas[3], Hassen Chaabani[4] and Pedro Moral*[,1]

[1]*Unitat d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain;* [2]*Centre d'Anthropologie, University Toulouse III, Toulouse, France;* [3]*Department of Biology, University of Crete, Herakleion, Greece;* [4]*Faculte de Pharmacie de Monastir, Universite du Centre, Monastir, Tunisia*

Alu elements are the most abundant mobile elements in the human genome (~1 100 000 copies). Polymorphic Alu elements have been proved to be useful in studies of human origins and relationships owing to two important advantages: identity by descent and absence of the Alu element known to be the ancestral state. Alu variation in the X chromosome has been described previously in human populations but, as far as we know, these elements have not been used in population relationship studies. Here, we describe the allele frequencies of 13 'young' Alu elements of the X chromosome (Ya5DP62, Ya5DP57, Yb8DP49, Ya5a2DP1, Yb8DP2, Ya5DP3, Ya5NBC37, Yd3JX437, Ya5DP77, Ya5NBC491, Yb8NBC578, Ya5DP4 and Ya5DP13) in six human populations from sub-Saharan Africa (the Ivory Coast), North Africa (Moroccan High Atlas, Siwa oasis in Egypt, Tunisia), Greece (Crete Island) and Spain (Basque Country). Eight out of 13 Alu elements have shown remarkably high gene diversity values in all groups (average heterozygosities: 0.342 in the Ivory Coast, 0.250 in North Africa, 0.209 in Europe). Genetic relationships agree with a geographical pattern of differentiation among populations, with some peculiar features observed in North Africans. Crete Island and the Basque Country show the lowest genetic distance (0.0163) meanwhile Tunisia, in spite of its geographical location, lies far from the other two North African samples. The results of our work demonstrate that X chromosome Alu elements comprise a reliable set of genetic markers useful to describe human population relationships for fine-scale geographical studies.
European Journal of Human Genetics (2007) **15,** 578–583. doi:10.1038/sj.ejhg.5201797; published online 28 February 2007

Keywords: Alu elements; X chromosome; human populations; Mediterranean region

Introduction

The X chromosome has the unique feature of being present only in a single copy in male subjects, which leads to a series of special characteristics that justify its increasing interest in studies of human population genetics. Mutations occur less frequently in this chromosome because every existing X chromosome spends two-thirds of its lifetime in female subjects where the nucleotide mutation rate is much lower than in male subjects. Additionally, the effective population size of the X chromosome is three quarters that of the autosomes. These facts contribute to the lower diversity of this chromosome (estimated to be about half of that on the autosomes). However, the smaller

*Correspondence: Professor P Moral, Department Biologia Animal-Antropologia, Faculty Biologia, Universidad de Barcelona, Avenue Diagonal 645, Barcelona 08028, Spain.
Tel: +34 934021461; Fax: +34 934035740;
E-mail: pmoral@ub.edu
Received 28 September 2006; revised 19 January 2007; accepted 24 January 2007; published online 28 February 2007

population size of the X chromosome also causes changes owing to genetic drift being faster than in other chromosomal regions and, therefore, population structure is expected to be more emphasized. Consequently, populations should differ more in their X chromosomes than in their autosomes. Linkage disequilibrium is also greater on the X chromosome, because only two-thirds of this chromosome manages to recombine in each generation. The size of regions with a single genetic history is expected to be larger than in autosomes, once more making it ideal for human population genetic studies. Finally, determining haplotypes on the X chromosome is a simple procedure and therefore it has proved to be a good choice for haplotype-based phylogenetic studies.[1]

X chromosome variation has been applied successfully in large-scale geographical studies that cope with the origin of non-African populations pointing in the direction of the 'Out of Africa' model.[2,3] However, the use of X chromosome for fine-scale geographical studies is negligible when compared with the high number of human population studies focused on Y chromosome and mtDNA variation.

Alu repeats are a category of short interspersed nuclear elements widely distributed in the genome of all primates. With a size of less than 500 bp, Alu elements are the most abundant mobile elements in the human genome ($\sim$1 100 000 copies). Twenty-five percent of the 'young' Alu human-specific elements have been incorporated in the human genome so recently that they are dimorphic for the presence or absence of the insertion. Subsequently, individuals can be polymorphic for the presence or absence of an Alu element at a particular chromosomal location. Alu elements have been shown to be useful in human evolution studies because they offer two important advantages compared with other polymorphic markers: (i) they are identical by descent – that is, individuals that share Alu-insertion polymorphisms have inherited them from a common ancestor and (ii) the ancestral state of each Alu insertion polymorphism is known to be the absence of the element, so that they can be used to 'polarize' the evolutionary process.[4]

The first thorough attempt to assess X chromosome variability relating to Alu elements was carried out in 2003 by Calliman *et al*.[5] The result was 345 Alu repeat elements from eight young Alu subfamilies, 264 of which were found on the X chromosome. From these elements, 16 were found to be polymorphic, with various levels of heterozygosity depending on the origin of each population. The heterozygosity data from that study suggested that the Alu elements from the X chromosome would be useful as genetic markers for human population genetics. The data showed a slight reduction in Alu polymorphism on the human sex chromosomes, in accordance with the observations made above. A brand new polymorphic Alu element, baptized DXS225 and belonging to the Ya5 Alu subfamily, was described by Pereira *et al*.[6] It was embedded in a LINE-1 retrotransposon, a region not previously examined, suggesting that there might actually be more polymorphic Alu elements on the X chromosome.

So far, the potential of Alu polymorphisms of the X chromosome, including a wide number of Alu polymorphisms of several anthropologically well-defined human populations, has not been tested. The Mediterranean region appears to be an interesting sample on which to test the efficacy of these markers. Previous anthropological works on Alu polymorphisms seem to be quite controversial, with some authors defending genetic flow between various groups in the Mediterranean and others assuming less interpopulation genetic relationships.[7–9] In this region, there are two levels of genetic flow that could be the subject of interesting surveys: among Mediterranean groups, and between Mediterranean and sub-Saharan groups.

In this study, 13 Alu elements scattered along the whole X chromosome were analyzed for the first time in six anthropologically well-defined groups selected by geographical, historical and ethnical criteria, including the Basques, several Mediterranean groups and a sub-Saharan sample from the Ivory Coast. The objectives were: (i) to explore the variation of polymorphic X chromosome Alu repeats in well-defined human populations; (ii) to apply this variation to the study of population relationships among South European and North African groups, and the various sub-Saharan influences in these populations and (iii) to provide new evidence of the usefulness of these markers to uncover the genetic variation between human populations.

Materials and methods

DNA samples

Blood samples were collected for DNA extraction and Alu amplification from six different populations (525 individuals), each one originating from a different country. Samples were obtained, with informed consent, from healthy and unrelated participants of both sexes. All participants had all four of their grandparents born in the same region. The work was approved by the Ethical Committee of the University of Barcelona. Two of the populations originated from anthropologically well-defined Berber groups in Morocco and Egypt (High Atlas and Siwa Oasis, respectively). The other samples were from Monastir (a Centre-North region of Tunisia), the Basque Country (in northern Spain) and Crete Island (Greece). Finally, a sample from the Ahizi ethnic group from the Ivory Coast was genotyped in order to include a representation of the sub-Saharan African variation.

Genetic determinations

PCR amplification was accomplished in 20-μl reactions for the 13 Alu sequences of the X chromosome (Ya5DP62, Ya5DP57, Yb8DP49, Ya5a2DP1, Yb8DP2, Ya5DP3, Ya5NBC37, Yd3JX437, Ya5DP77, Ya5NBC491, Yb8NBC578, Ya5DP4 and Ya5DP13). Primer sequences are described in Calliman

et al,[5] as well as amplification and electrophoresis conditions with minor modifications.

Statistical analysis

Allele frequencies were calculated by direct gene counting. χ^2 and Fisher's Exact tests were used to detect significant differences between male and female allele frequencies, and to check the Hardy–Weinberg equilibrium for each locus[10] using the GENEPOP v3.3[11] statistical package. Reynolds' distances[12] were calculated between pairs of populations with PHYLIP[13] statistical package and the consistency of the distance values was checked by bootstrap resampling analysis (100 iterations). Principal component analysis (PCA) was carried out using the R-MATRIX program.[14] Polymorphism and population structure within and between groups was tested by analysis of molecular variance (AMOVA) using the ARLEQUIN v2.0 program.[15] Global F_{ST} values were estimated by averaging partial values, and the resultant probability was calculated by combining probabilities from each individual test.[16]

Apart from the usual distance-based clustering methods described above, a model-based method was also used to infer population structure by means of the STRUCTURE 2.1 program.[17] A model of K population groups (where K might be unknown) was assumed. This model was tested for several values of K using a specific Markov Chain Monte Carlo algorithm (the Gibbs sampler). STRUCTURE estimates the 'natural logarithm of the probability of the data' for each value of K, briefly referred to as 'Ln P(X\K)'. Among the estimated K values, that yielding the lowest absolute value of the Ln P(X\K) is the one that best describes the data. In our data sets, we ran the Gibbs sampler under the admixture model (INFERALPHA = 1.0), using prior population information and assuming correlated allele frequencies. All runs included a burn-in period of 50 000 iterations followed by 10^6 iterations, and they were repeated three times each in order to test the consistency of the results.

Results

Allele frequencies, Hardy–Weinberg equilibrium and heterozygosity

Table 1 shows the frequencies of the 13 Alu insertion alleles (Alu+) for the six populations. Some Alu elements appear to be fixed in several populations. In most cases, the frequency distributions fitted the Hardy–Weinberg equilibrium. The raw deviations found in Ya5DP62 in Ivory Coast ($P = 0.003$), Ya5DP57 in Basques ($P = 0.007$) and Ya5NBC491 in Tunisia ($P = 0.043$) were not significant after the Bonferroni correction. Heterozygosity levels per locus and per population are also shown in Table 1. In general, most Alu elements show moderate to high diversity, except for Ya5NBC491, Yb8NBC578, Ya5DP4 and Ya5DP13. The Ya5NBC37 Alu shows the highest

Table 1 X chromosome Alu insertion frequencies and heterozygosity per locus and per population

	Ivory Coast	*High Atlas*	*Siwa Oasis*	*Tunisia*	*Crete Island*	*Basque Country*	*H per locus (Ivory Coast)*	*H per locus (rest)*
Ya5DP62	(82) 0.634	(147) 0.871	(123) 0.821	(168) 0.744	(116) 0.802	(87) 0.713	0.464	0.325
Ya5DP57	(84) 0.262	(140) 0.636	(137) 0.431	(159) 0.830	(86) 0.744	(115) 0.774	0.387	0.393
Yb8DP49	(85) 0.706	(149) 0.779	(138) 0.804	(168) 0.744	(120) 0.842	(136) 0.728	0.415	0.341
Ya5a2DP1	(85) 0.294	(151) 0.669	(140) 0.629	(159) 0.830	(119) 0.874	(136) 0.869	0.415	0.328
Yb8DP2	(83) 0.711	(151) 0.305	(142) 0.359	(165) 0.200	(121) 0.231	(138) 0.130	0.411	0.357
Ya5DP3	(87) 0.230	(151) 0.199	(142) 0.239	(164) 0.232	(120) 0.067	(134) 0.134	0.354	0.279
Ya5NBC37	(80) 0.313	(146) 0.432	(142) 0.275	(170) 0.412	(121) 0.339	(136) 0.324	0.430	0.452
Yd3JX437	(67) 0.552	(150) 0.213	(138) 0.246	(152) 0.033	(98) 0.143	(125) 0.104	0.495	0.240
Ya5DP77	(84) 0.286	(136) 0.912	(128) 0.914	(166) 0.892	(109) 0.982	(90) 1.000	0.400	0.093
Ya5NBC491	(67) 0.940	(148) 1.000	(143) 0.993	(158) 0.975	(121) 1.000	(132) 1.000	0.124	0.055
Yb8NBC578	(71) 0.789	(150) 0.940	(139) 0.950	(165) 0.933	(121) 1.000	(133) 0.932	0.333	0.092
Ya5DP4	(73) 0.000	(150) 0.013	(143) 0.000	(165) 0.012	(119) 0.059	(133) 0.067	0.000	0.047
Ya5DP13	(71) 0.901	(149) 0.993	(143) 1.000	(167) 0.988	(121) 1.000	(133) 1.000	0.195	0.009
H per population	0.3423 (0.1461)	0.2530 (0.1818)	0.2653 (0.1855)	0.2305 (0.1575)	0.2066 (0.1767)	0.2118 (0.1656)		

The parentheses before the frequencies show the number of chromosomes examined. The parentheses after the heterozygosities in the last row indicate the standard deviation. The heterozygosity per locus (two last columns) is indicated separately for the Ivory Coast and for the rest of the samples, owing to important heterozygosity differences between these two groups.

Table 2 Reynolds' distances (below diagonal) and respective distance errors (above diagonal) for the six populations studied

	Ivory Coast	*Crete Island*	*Basque Country*	*Siwa Oasis*	*High Atlas*	*Tunisia*
Ivory Coast		0.0718	0.0768	0.0633	0.0564	0.0660
Crete Island	0.314490		0.0536	0.0255	0.0107	0.0089
Basque Country	0.322052	0.016299*		0.0265	0.0099	0.0055
Siwa Oasis	0.176117	0.072466	0.086411		0.0104	0.0347
High Atlas	0.215017	0.035995	0.048536	0.022322		0.0131
Tunisia	0.297528	0.029565	0.016804	0.086113	0.038674	

Only the distance between Basque Country and Crete Island was found not significantly different from zero marked with an asterisk.

heterozygosity (0.436), followed by Ya5DP57, Yb8DP49, Ya5a2DP1 and Yb8DP2 Alu elements. Likewise, the most diverse population seems to be the Ivory Coast ($H=0.342$) followed by the rest of the North African (average $H=0.250$) and South European populations (average $H=0.209$). No significant linkage disequilibrium was present in any pair of the Alu markers, in accordance with the large chromosomal distances between them (ranging from 560 Kb to 59.65 Mb).

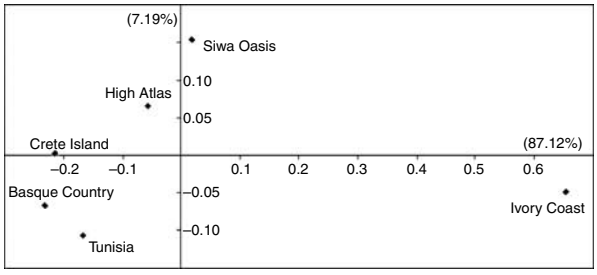


Figure 1 Principal components analysis of the Ivory Coast, Basque Country, Crete Island, High Atlas, Siwa Oasis and Tunisia. The first two axes account for 94.31% of the total variance. The percentage of variance accounted for by each axis is shown in parentheses.

Reynolds' distances and principal components analysis

Population pairwise comparisons are indicated in Table 2. Bootstrap resampling analysis for the calculation of distance errors showed high consistency of values. The lowest distance has been found between Crete Island and Basque Country (0.0163), and this is the only value not significantly different from zero. On average, the mean distance of Ivory Coast to the other five samples is 0.2650, whereas the mean distance among Mediterraneans and Basques is considerably lower (0.0453).

The PC analysis allowed the graphic representation of population relationships, as shown in Figure 1. The first two axes account for 94.31% of the total variance when the six samples are considered. The first axis clearly separates the sub-Saharan group of the Ivory Coast from the rest. The two European samples lie in the other extreme of the variation, whereas the three North African samples show an intermediate position, although always closer to the Europeans. The second component contributes to a relative separation among non-sub-Saharan groups, with Siwa Oasis and Tunisia appearing at the most distant positions. When the analysis was repeated, in order to remove the effect of the sub-Saharan sample (not shown), the relationship pattern among the remaining populations was substantially the same. In this case, the first axis (accounting for 64.79% of the total variance) underlines the separation of Siwa Oasis and High Atlas from the Basques, Crete Island and Tunisia. It is interesting to note that Tunisia is the North African group that lies genetically closest to the European samples.

Analysis of molecular variance

A first approach to population diversity through the F_{ST} statistic yields a global value of 9.88% ($P<0.001$) when all six populations were grouped together (data not shown). That is, almost 10% of the detected genetic variance was concentrated between populations. By locus, the F_{ST} values range from 0.99% ($P<0.05$, for Ya5NBC37) to an impressive 40.38% ($P<0.0001$, for Ya5DP77). The global F_{ST} (9.88) is mainly attributed to, on the one hand, the variation of Ya5DP57, Ya5a2DP1, Yb8DP2, Yd3JX437 and Ya5DP77 Alu markers, and, on the other hand, the inclusion of the Ivory Coast sample. By excluding the Ivory Coast, the global F_{ST} changed to a more moderate value (3.81%, $P<0.001$). This mainly reflects the variation of the Ya5DP57 Alu polymorphism (individual $F_{ST}=11.50\%$, $P<0.0001$), which in Siwa Oasis presents clearly distinguishable allele frequencies.

A hierarchical AMOVA, assuming two geographical groups ('South European' and 'North African') without the Ivory Coast, does not reveal any significant differences between the two Mediterranean shores, the frequency variance between the two groups ($F_{CT}=1.24\%$, $P=0.338$) being clearly lower than the diversity among populations within groups ($F_{SC}=3.06\%$, $P<0.001$). Separate AMOVA tests for South European and North African samples gave F_{ST} values of 0.38 and 3.45, respectively, both of them significant ($P<0.001$).

Table 3 Estimated natural logarithm of the probability of the data (Ln P(X\K)) for each value of K in three data sets and proportion of membership of each pre-defined population in each of the three clusters

K	*Ln P(X\K)*		
	All populations	*All Mediterranean populations*	
1	−4178.6	−3404.0	
2	−4043.4	−3531.9	
3	−4007.2	−3528.4	
4	−4094.6	−3723.2	
5	−4317.1	−3937.1	
6	−4450.8	—	
	Cluster 1	*Cluster 2*	*Cluster 3*
Ivory Coast	0.770	0.126	0.104
Crete Island	0.194	0.379	0.427
Basque Country	0.188	0.422	0.390
Tunisia	0.208	0.397	0.394
Siwa Oasis	0.388	0.300	0.312
High Atlas	0.323	0.324	0.353

Model-based inference of population structure

To explore the degree of genetic structure among our samples, we defined a STRUCTURE data file containing all six populations. We estimated the posterior probabilities departing from the Ln P(X\K) value. As seen in Table 3, when surveying population structure in all six populations, the model with $K=3$ seems to fit our data best. The same table also indicates the proportion of membership of each pre-defined population in each of the three clusters inferred for data set. The pattern of membership of the Ivory Coast is the most differentiated, with 77% of membership in cluster 1. The European and Tunisian samples show a similar pattern of membership in the three clusters (1:2:2), whereas Siwa Oasis and High Atlas seem to follow a 1:1:1 pattern. As expected, no cluster gets to become exclusively characteristic of the populations implicated, although cluster 1 could be considered as representative of the sub-Saharan variation. To delve into the two clusters not directly related with the sub-Saharan variation, we repeated the runs including only the five non-sub-Saharan groups (see Table 3). This time the model with $K=1$ was best to describe the data.

Discussion

This paper describes the pattern of the frequency distribution of 13 polymorphic Alu insertions of the X chromosome in five well-defined groups from the Mediterranean region, including the Basque Country. In general, the allele frequencies found range within the general patterns described previously,[5] but with a remarkable between-population variation. In terms of variation, the six samples were adjusted to a decreasing pattern of diversity from South to North (mean heterozygosity for the Ivory Coast: 0.342, for North African samples: 0.250 and for the European ones: 0.209).

The hierarchical AMOVA analysis in the five Mediterranean groups (Basques included) showed that only a small and nonsignificant part of the genetic variance could be attributed to the variation between North–South groups ($F_{CT}=1.24\%$), indicating no particular genetic differentiation between the two sides of the Mediterranean Sea. However, the markers examined are consistent with a more important diversity within North Africa ($F_{ST}=3.45\%$, $P<0.001$) than in South Europe (0.38%, $P<0.001$). Comas *et al*[7] and González-Pérez *et al*[8] had previously studied population relationships in the western-Mediterranean basin using polymorphic autosomal Alu elements. These two surveys indicated a North *versus* South differentiation (F_{CT} values: 1.80 and 1.96%, respectively) slightly higher than in our results, but a population variation within groups clearly lower (F_{SC} values of 2.30 and 0.47%, respectively) than that evidenced from the X chromosome Alu markers (F_{SC} of 3.06%.). Moreover, the 13 Alu markers of this study reveal a population variation within North Africa ($F_{ST}=3.45\%$) sixfold higher than that obtained from a similar number of autosomal Alu elements ($F_{ST}=0.57\%$; Comas *et al*, 2000). The X chromosome markers reveal a higher population differentiation in comparison with the same kind of genetic markers in autosomes. This possibly reflects the effect of the reduced population size of the X chromosome on population variation, which, as we mentioned in the introduction, makes populations differ more in their X chromosomes than in autosomal markers.

The Reynolds' distances revealed generally accepted relationships among the Mediterranean populations. It is remarkable that the Arab-speaking sample from Tunisia shows a particular genetic position as compared with other North African groups. In fact, the Tunisian genetic distances to European samples are smaller than those to North African groups. This close position of Tunisia to the Europeans also appears in the population distribution in the PCA graph (Figure 1). This could be explained by the history of the Tunisian population, reflecting the influence of the ancient Phoenician settlers of Carthage followed, among others, by Roman, Byzantine, Arab and French occupations, according to historical records. Notwithstanding, other explanations cannot be discarded, such as the relative heterogeneity within current Tunisian populations,[18] and/or the limited sub-Saharan genetic influence in this region as compared with other North African areas, without excluding the possibility of the genetic drift, whose effect might be particularly amplified on the X chromosome.

An interesting aspect comes from the evidenced relationships between the Basque Country and Crete Island. These two populations have distinct historical, anthropological and cultural backgrounds, and yet no significant differ-

ences were found between them when a locus-by-locus χ^2 comparison was carried out. As for the remaining analyzed populations, Siwa Oasis seems to be the most differentiated (see Table 2 and Figure 1). The differentiation shown by Siwa Oasis, and also by High Atlas, could be related to higher foreign genetic contributions, from West Sahara into High Atlas and Nile groups into the Siwa Oasis. Esteban *et al*[19] described a similar pattern of GGC allele frequencies of the androgen receptor (located in chromosome X) for the Ivory Coast and Siwa Oasis samples, giving evidence of sub-Saharan genetic influence in this Berber group.

The model-based method indicated population structure with three clusters inferred when all populations were examined. Cluster 1 (Table 3) is evidently the 'sub-Saharan' one. As for the other two, they seem to have a biological meaning only when seen in comparison with cluster 1. All European and North African samples show a 1:1 membership proportion in clusters 2 and 3. Only when seen in conjunction with cluster 1 do differences appear. The Basque country, Crete Island and Tunisia feature a similar pattern of membership in the three clusters (1:2:2), whereas Siwa Oasis and High Atlas seem to follow a 1:1:1 pattern. This could possibly explain why the software has failed to detect population structure in the Mediterranean populations, implying that there are no population-specific genetic patterns representative enough to allow us to assign, with certainty, individuals to populations.

The failure to detect population structure among Mediterranean groups might lead to the conclusion that our data are controversial, because we have already discussed the striking differences of, for example, Siwa Berbers and High Atlas Berbers from the other groups. Apparently, differences do exist, but they are not striking enough to allow the definition of different clusters within the Mediterranean region. Clustering appears only when a quite distinct human group is added, such as the sub-Saharan Ahizi from the Ivory Coast.

To sum up, our data on X chromosome markers support, in general, the differentiation patterns of the Mediterranean populations described by other investigators, providing, at the same time, detailed data of the frequency distribution of X chromosome Alu elements. To our knowledge, it is the first time that these specific molecular markers have been used in such a study. X chromosome Alu elements seem to perform well in fine-scale population differentiation studies. As no completely comparable data exist to survey the effect of the reduced population size of the X chromosome on the genetic distances between populations, a future investigatory line should include data from autosomal markers from the same populations. Furthermore, linkage disequilibrium studies are advisable, including higher mutation rate STR markers close to the X chromosome Alu insertions.

Acknowledgements

We are grateful to all of the donors for providing blood samples and the people who contributed to the collection. In particular, we thank Professor André Chaventré and Dr Gil Bellis (for the samples from Ivory Coast). This research has been supported by the Ministerio de Ciencia y Tecnología CGL 2005-3391 and Generalitat de Catalunya SGR00252 projects. The sampling of the Berbers from Morocco and Egypt was supported by the Conseil Régional de Midi-Pyrénées, Toulouse (France). The work of GA has been financed by an FPU grant from the Ministerio de Educación y Ciencia (grant reference: AP2005-4425).

References

1 Schaffner SF: The X chromosome in population genetics. *Nat Rev Genet* 2004; **5**: 43–51.
2 Harris EE, Hey J: X chromosome evidence for ancient human histories. *Proc Natl Acad Sci U S A* 1999; **96**: 3320–3324.
3 Yu N, Fu YX, Li WH: DNA polymorphism in a worldwide sample of human X chromosomes. *Mol Biol Evol* 2002; **19**: 2131–2141.
4 Batzer MA, Deininger PL: Alu repeats and human genomic diversity. *Nat Rev Genet* 2002; **3**: 370–379.
5 Calliman PA, Hedges DJ, Salem AH *et al*: Comprehensive analysis of Alu-associated diversity on the human sex chromosomes. *Gene* 2003; **317**: 103–110.
6 Pereira RW, Santos SS, Pena SD *et al*: A novel polymorphic Alu insertion embedded in a LINE 1 retrotransposon in the human X chromosome DXS225: identification and worldwide population study. *Genet Mol Res* 2006; **5**: 63–71.
7 Comas D, Calafell F, Benchemsi N *et al*: Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Hum Genet* 2000; **107**: 312–319.
8 Gonzalez-Perez E, Via M, Esteban E *et al*: Alu insertions in the Iberian Peninsula and north west Africa – genetic boundaries or melting pot? *Coll Antropol* 2003; **27**: 491–500.
9 Garcia-Obregon S, Alfonso-Sanchez MA, Perez-Miranda AM *et al*: Genetic position of Valencia (Spain) in the Mediterranean basin according to Alu insertions. *Am J Hum Biol* 2006; **18**: 187–195.
10 Guo SW, Thompson EA: Performing the exact test of Hardy–Weinberg proportions for multiple alleles. *Biometrics* 1992; **48**: 361–372.
11 Raymond M, Rousset F: GENEPOP version 1.2 population genetics software for exact tests and ecumenicism. *J Hered* 1995; **86**: 248–249.
12 Reynolds J, Weir BS, Cockerman CC: Estimation of the coancestry coefficient: basis for a short term genetic distance. *Genetics* 1983; **105**: 767–779.
13 Felsenstein J: PHYLIP – phylogeny inference package. *Cladistics* 1989; **5**: 164–166.
14 Harpending H, Jenkins T: Genetic distance among southern African populations; in Crawford MH, Workman PL (eds): *Methods and Theories of Anthropological Genetics*. University of New Mexico Press, 1973, pp 177–199.
15 Schneider S, Roessli D, Excoffier L: Arlequin: A software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, Geneva, 2000.
16 Sokal RR, Rohlf FJ: *Biometry*. Second edition, Freeman and Co, New York, 1981.
17 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–959.
18 Giraldo MP, Esteban E, Aluja MP *et al*: Gm and Km alleles in two Spanish Pyrenean populations (Andorra and Pallars Sobira): a review of Gm variation in the Western Mediterranean basin. *Ann Hum Genet* 2001; **65**: 537–548.
19 Esteban E, Rodon N, Via M *et al*: Androgen receptor CAG and GGC polymorphisms in Mediterraneans: repeat dynamics and population relationships. *J Hum Genet* 2006; **51**: 129–136.

Results II

Athanasiadis et al., 2009

Polymorphism FXII 46C>T and cardiovascular risk: additional data from Spanish and Tunisian patients

Georgios Athanasiadis, Esther Esteban, Magdanela Gayà-Vidal, Robert Carreras-Torres, Raoudha Bahri and Pedro Moral

BMC Research Notes 2009; 2: 154

Resumen en castellano

El gen que codifica para el Factor de Coagulación XII (F12) ha sido objeto de muchos estudios epidemiológicos de riesgo cardiovascular, con resultados controvertidos.

Estudios previos han mostrado que los niveles del FXII en el plasma de la sangre exhiben un 67% de heredabilidad y que dichos niveles son probablemente determinados exclusivamente por variantes dentro del gen F12. De estas variantes, la más estudiada es la transición 46C>T. Se ha mostrado que hay una asociación entre este polimorfismo y la variación en los niveles del factor XII en el plasma de la sangre, ya que se cree que el 46C>T afecta la eficiencia de la traducción. Asimismo, estudios de caso-control en muestras españolas indicaron que el genotipo T/T de dicho polimorfismo es un independiente factor de riesgo para el desarrollo de trombosis venosa, ataque cerebral isquémico y enfermedad coronaria aguda.

Este estudio intenta confirmar la importancia del polimorfismo 46C>T para el riesgo cardiovascular utilizando una muestra de Barcelona y otra de Túnez y un

doble diseño epidemiológico: por un lado, se realizó un análisis de asociación basado en familias mediante el desequilibrio de transmisión (Transmission Disequilibrium Test – TDT) en 101 familias nucleares de Barcelona con un hijo afectado por cardiopatía isquémica y, por otro, se llevó a cabo un estudio clásico de caso-control basado en 76 pacientes con cardiopatía isquémica y en 119 controles sanos del norte y centro-sur de Túnez.

Los sujetos fueron genotipados para el polimorfismo 46C>T con una prueba TaqMan de Applied Biosystems y los datos fueron analizados para investigar potenciales asociaciones entre la variante T y la manifestación de la enfermedad.

Los análisis estadísticos no revelaron asociación en ninguna de las dos muestras (estudio de familias: riesgo relativo=1.17, $p>0.05$; estudio caso-control: odds ratio=1.36, $p>0.05$). En el caso del estudio caso control, dada la alta frecuencia del alelo T, había suficiente poder estadístico para detectar un riesgo del mismo orden de magnitud que el descrito en la bibliografía (4.8), mientras que el estudio de familias carecía suficiente poder para detectar un riesgo relativo tan bajo como el descrito en este trabajo (mínimo riesgo relativo detectable por nuestra muestra: 1.90).

Nuestros resultados sugieren que el polimorfismo 46C>T no supone un factor de riesgo de cardiopatía isquémica en ninguna de las dos muestras analizadas, aunque la incorporación de más familias nucleares sería deseada. En el contexto de las controversias existentes en la literatura, nuestros datos son más compatibles con la tesis expresada por otros según la cual los niveles del FXII en la sangre probablemente no sean una causa sino una consecuencia de la trombosis.

Supervisor's report on the involvement of the PhD student in the development of this paper

UNIVERSITAT DE BARCELONA

Dr. **Pedro Moral Castrillo**, Professor at the Department of Animal Biology of the University of Barcelona and supervisor of the doctoral thesis "Genetic variation of the X chromosome and the genomic regions of Coagulation Factors VII and XII in human populations: Epidemiological and evolutionary considerations" by **Georgios Athanasiadis**, hereby certifies that the participation of the above student in the article "**Polymorphism FXII 46C>T and cardiovascular risk: additional data from Spanish and Tunisian patients**", published in *BMC Research Notes,* consisted of the following tasks:

- Participation in the design of the study together with Dr. Pedro Moral Castrillo

- Genotype determination of the 46C>T polymorphism in the lab

- Creation of the genotype database and statistical analysis of the data

- Participation in the manuscript drafting together with Dr. Pedro Moral Castrillo

In addition, it is important to note that none of the co-authors of this article have used the results of this work in any implicit or explicit way to develop another doctoral thesis. As a consequence, this article forms part of the doctoral thesis of Georgios Athanasiadis exclusively.

Signed by Dr. Pedro Moral Castrillo

Barcelona, 9 April 2010

BMC Research Notes

Short Report

Polymorphism FXII 46C>T and cardiovascular risk: additional data from Spanish and Tunisian patients

Georgios Athanasiadis[1], Esther Esteban[1], Magdanela Gayà Vidal[1], Robert Carreras Torres[1], Raoudha Bahri[2] and Pedro Moral*[1]

Address: [1]Unitat d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain and [2]Faculté de Pharmacie de Monastir, Université de Monastir, Monastir, Tunisia

Email: Georgios Athanasiadis - athanasiadis@ub.edu; Esther Esteban - mesteban@ub.edu; Magdanela Gayà Vidal - magdagaya@ub.edu; Robert Carreras Torres - robertcarrerastorres@hotmail.com; Raoudha Bahri - bahri@yahoo.fr; Pedro Moral* - pmoral@ub.edu

* Corresponding author

Published: 31 July 2009

BMC Research Notes 2009, **2**:154 doi:10.1186/1756-0500-2-154

Received: 30 May 2009
Accepted: 31 July 2009

This article is available from: http://www.biomedcentral.com/1756-0500/2/154

Abstract

Background: Previous studies showed an association between Coagulation Factor XII 46C>T polymorphism and variation in FXII plasma levels, as 46C>T seems to affect the translation efficiency. Case-control studies in Spanish samples indicated that genotype T/T is an independent risk factor for venous thrombosis, ischemic stroke and acute coronary artery disease. In this study, we tried to reaffirm the importance of 46C>T in two samples from Spain and Tunisia.

Findings: A Transmission Disequilibrium Test (TDT) based on 101 family trios from Barcelona with one offspring affected by ischemic heart disease and a classical case-control study based on 76 patients with IHD and 118 healthy individuals from North and Centre-South Tunisia were conducted. Subjects were genotyped for 46C>T and data were analyzed accordingly, revealing no association in any of the two samples (TDT: P = 0.16, relative risk 1.17; case-control study: P = 0.59, odds ratio 1.36).

Conclusion: The results suggest that 46C>T is not a risk factor for ischemic heart disease in any of the two analyzed samples and therefore the polymorphism seems not to be a universal risk factor for cardiovascular diseases.

Findings

The gene of coagulation Factor (F) XVII has been repeatedly analysed in many epidemiological studies on cardiovascular (CV) genetic risk with controversial results. It has been reported that FXII plasma levels exhibit a 67% of heritability and that these levels are strongly determined by F12 gene variants.[1] Although low FXII plasma levels have been reported to affect the development of CV diseases, it has been recently proposed that, rather than the cause of thrombosis, these low levels are actually the result of it. [2,3]

A highlighted F12 polymorphism is the 46C>T transition. It seems that 46C>T affects the translation efficiency, leading to reduced FXII plasma levels.[4] This polymorphism (with allele frequencies 0.8/0.2 in Caucasians) accounts for 40% of the variance in FVII activity in a Spanish Mediterranean population and seems to fit a log additive

model of inheritance with an allele dose-dependent effect.[1,4]

Previous genetic association studies between 46C>T and CV risk showed that genotype T/T increases significantly the risk for venous thrombosis, ischemic stroke and acute coronary artery disease in a Spanish population. [5-7] The odds ratio in acute coronary artery disease was estimated around 4.8.[7] Conversely, there are also studies that either do not detect genetic association between this genotype and CV risk or alternatively report association between the C/C genotype and CV risk.[2,8]

In this context, this study provides additional data about the allele frequency distribution of 46C>T in two new Mediterranean samples. These data are used in an attempt to confirm previous genetic association studies, by a double epidemiological design: a family-based analysis (Transmission Disequilibrium Test TDT) applied in a Spanish sample and a classical case-control study applied in a Tunisian sample.

A total of 101 nuclear families (n = 302) with one offspring (93 males and 8 females) clinically diagnosed for ischemic heart disease (IHD) were analyzed, all from the area of Barcelona, Spain. A comprehensive description of the families, diagnosis criteria and DNA extraction methods has been reported elsewhere.[9] Moreover, 76 patients with IHD complicated by myocardial infarction, all from Monastir, Tunisia, as well as 118 unrelated healthy autochthonous individuals from North and Central-North Tunisia were also analyzed in an unmatched case-control design.[10] This study has been performed in accordance with the Ethical Committee guidelines of the participating Hospitals and Universities and all subjects participating in the study signed a written informed consent.

Polymorphism 46C>T was genotyped by Real-Time PCR, using the standard TaqMan® SNP genotyping assay protocol of Applied Biosystems for a total volume of 5 μl per well. DNA amplification as well as fluorescence intensity measurements of the final reaction product and data collection were carried out with an ABI PRISM® 7900 HT Sequence Detection System.

Allele frequencies in the Spanish IHD patients and their parents as well as in patients and controls from Tunisia were calculated directly. A TDT analysis was performed in the IHD families from Barcelona, using the FBAT v2.0.2c program.[11] In the Tunisian case-control study, association was examined with Epi Info™ v3.4.3.[12] This program was also used for the calculation of the relative risk in the TDT analysis and the odds ratio (T/T vs. C/C and C/T) in the case-control design. Sample power was calculated using QUANTO v1.2.3 assuming a CV disease prevalence of 0.07.[13]

Frequencies of the T allele in the Spanish and Tunisian samples are shown in Table 1. In the Spanish families, the TDT analysis indicated a low relative risk of 1.17 (95% CI: 0.75 <RR<1.82) for the T allele, without any significant transmission to the patients (p = 0.16). Power calculations revealed that the minimum detectable risk our family sample size (n = 101) could find was approximately 1.9 (with a power >80%). In the Tunisian sample, the estimated odds ratio for the T/T genotype was 1.36 (95% CI: 0.36 < OR <4.93), without any significant association (p = 0.59). Again, the minimum detectable effect our set of 76 patients and 118 controls could find was around 1.9. Finally, no significant differences were found in any pairwise comparison of the data in Table 1.

Our study showed that the allele frequencies of the FXII 46C>T polymorphism are similar in the Spanish and Tunisian samples, in accordance with the Caucasian pattern. This suggests a stable presence of 46C>T in different Mediterranean regions. On the other hand, our results indicate the absence of association between 46C>T and CV disease in both samples.

The similar allele frequency distributions between Spanish patients and parents and between Tunisian patients and controls provide a first rough estimate of the lack of association. Moreover, our Tunisian sample has enough power (>99%) to detect an odds ratio of 4.8, as reported

Table 1: Allele and genotype frequencies of the 46C>T in the Spanish and Tunisian CV patients

		Allele frequencies		Genotype frequencies			
		C	T	CC	CT	TT	N
Spain	affected children	0.768	0.232	0.590	0.356	0.054	97
	parents	0.799	0.201	0.638	0.321	0.040	137
Tunisia	cases	0.770	0.230	0.593	0.354	0.053	76
	controls	0.750	0.250	0.563	0.375	0.063	118

elsewhere, reinforcing the consistency of the lack of association detected.[7] As far as we know, this is the first time the potential role of 46C>T is tested through a family-based TDT analysis. Out findings confirm the negative results of the case-control design, always taking into account that our family trios have power only to detect risks above 1.9.

The diverse results of previous studies regarding association between polymorphism 46C>T and CV disease could be explained, at least in part, by the differences in study design, the definition of inclusion and exclusion criteria and the number of participants. In our case, the low sample size in the TDT study and the lack of a matched case-control design impose some caution in the interpretation of our results.

In short, our results provide indirect evidence that 46C>T is not a universal independent risk factor for CV diseases. What seems to be well established is that the T allele affects the variation of intermediate phenotypes (FXII plasma levels). If this is true, then the diversity of results in genetic association studies is hard to explain. This observation leads us to adopt the idea that the low levels of FXII are the consequence rather than the cause of the disease, as has been recently suggested.[3] This idea finds solid ground when combined with other independent results on the lack of causality between FXII deficiency and venous thrombosis. [14,15] In this context, 46C>T polymorphism may not be a direct determinant of CV risk.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GA carried out the SNP genotyping, participated in the statistical analyses and wrote the original manuscript. EE gave important advice for the improvement of the manuscript. MGV carried out the Spanish sample preparation. RCT helped with the statistical analyses. RB carried out the DNA extraction from the Tunisian samples. PM participated in the design and coordination of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This research has been supported by the Ministerio de Ciencia y Tecnología CGL2005-3391 and Generalitat de Catalunya SGR00252 projects. The work of GA has been financed by an FPU grant from the Ministerio de Ciencia e Innovación (grant reference: AP2005-4425).

References

1. Soria JM, Almasy L, Souto JC, Bacq D, Buil A, Faure A, Martínez-Marchán E, Mateo J, Borrell M, Stone W, Lathrop M, Fontcuberta J, Blangero J: **A quantitative trait locus in human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease.** *Am J Hum Genet* 2002, **70:**567-574.
2. Bach J, Endler G, Winkelmann BR, Boehm BO, Maerz W, Mannhalter C, Hellstern P: **Coagulation factor XII activity, activated factor XII, distribution of factor XII C46T gene polymorphism and coronary risk.** *J Thromb Haemost* 2008, **6:**291-296.
3. Kanaji T: **Lower factor XII activity is a risk marker rather than a risk factor for cardiovascular disease: a rebuttal.** *J Thromb Haemost* 2008, **6:**1053-1054.
4. Kanaji T, Okamura T, Osaki K, Kuroiwa M, Shimoda K, Hamasaki N, Niho Y: **A common genetic polymorphism (46 C to T substitution) in the 5'-untranslated region of the coagulation factor XII gene is associated with low translation efficiency and decrease in plasma factor XII level.** *Blood* 1998, **91:**2010-2014.
5. Tirado I, Soria JM, Mateo J, Oliver A, Souto JC, Santamaria A, Felices R, Borrell M, Fontcuberta J: **Association after linkage analysis indicates that homozygosity for the 46C-->T polymorphism in the F12 gene is a genetic risk factor for venous thrombosis.** *Thromb Haemost* 2004, **92:**892-893.
6. Santamaría A, Mateo J, Tirado I, Oliver A, Belvís R, Martí-Fábregas J, Felices R, Soria JM, Souto JC, Fontcuberta J: **Homozygosity of the T allele of the 46 C->T polymorphism in the F12 gene is a risk factor for ischemic stroke in the Spanish population.** *Stroke* 2004, **35:**1795-1799.
7. Santamaría A, Martínez-Rubio A, Mateo J, Tirado I, Soria JM, Fontcuberta J: **Homozygosity of the T allele of the 46 C-->T polymorphism in the F12 gene is a risk factor for acute coronary artery disease in the Spanish population.** *Haematologica* 2004, **89:**878-879.
8. Kanaji T, Watanabe K, Hattori S, Urata M, Iida H, Kinoshita S, Kayamori Y, Kang D, Hamasaki N: **Factor XII gene (F12) -4C/C polymorphism in combination with low protein S activity is associated with deep vein thrombosis.** *Thromb Haemost* 2006, **96:**854-855.
9. Via M, López-Alomar A, Valveny N, González-Pérez E, Bao M, Esteban E, Pintó X, Domingo E, Moral P: **Lack of association between eNOS gene polymorphisms and ischemic heart disease in the Spanish population.** *Am J Med Genet A* 2003, **116:**243-248.
10. Bahri R, Esteban E, Moral P, Chaabani H: **New insights into the genetic history of Tunisians: Data from Alu insertion and apolipoprotein E gene polymorphisms.** *Ann Hum Biol* 2008, **35:**22-33.
11. Laird N, Horvath S, Xu X: **Implementing a unified approach to family based tests of association.** *Genet Epidemiol* 2000, **19:**36-42.
12. Dean AG, Dean JA, Burton AH, Dicker RC: **Epi Info™: a general purpose microcomputer program for health information systems.** *Am J Preventive Medicine* 1991, **7:**178-182.
13. Gauderman WJ, Morrison JM: **QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies.** 2006 [http://hydra.usc.edu/gxe].
14. Girolami A, Randi ML, Gavasso S, Lombardi AM, Spiezia F: **The occasional venous thromboses seen in patients with severe (homozygous) FXII deficiency are probably due to associated risk factors: a study of prevalence in 21 patients and review of the literature.** *J Thromb Thrombolysis* 2004, **17:**139-143.
15. Lombardi AM, Bortoletto E, Scarparo P, Scapin M, Santarossa L, Girolami A: **Genetic study in patients with factor XII deficiency: a report of three new mutations exon 13 (Q501STOP), exon 14 (P547L) and -13C>T promoter region in three compound heterozygotes.** *Blood Coagul Fibrinolysis* 2008, **19:**639-643.

Results III

Athanasiadis et al., 2010a

Different evolutionary histories of the coagulation factor VII gene in human populations?

Georgios Athanasiadis, Esther Esteban, Magdalena Gayà-Vidal, Jean-Michel Dugoujon, Nicholas Moschonas, Hassen Chaabani, Nisrine Bissar-Tadmouri, Nourdin Harich, Mark Stoneking and Pedro Moral

Annals of Human Genetics 2010; 74(1): 34-45

Resumen en castellano

Las anomalías en la coagulación de la sangre constituyen un factor de riesgo para enfermedades cardiovasculares en sociedades industrializadas. Sin embargo, es posible que una mayor capacidad de coagulación haya aportado una ventaja en la supervivencia de nuestros antepasados en forma de una recuperación más rápida tras una hemorragia postraumática o posparto.

Este trabajo investiga la historia evolutiva del gen del Factor de Coagulación VII (F7) en 16 poblaciones procedentes de la región mediterránea más amplia, de África subsahariana y de Bolivia.

Para ello, se han analizado 5 mutaciones del promotor del gen F7 asociadas con el riesgo cardiovascular (-670A>C, -630A>G, -402G>A, -401G>T y -324 ins10bp) y 9 polimorfismos neutros (6 polimorfismos de un solo nucleótido – SNPs – y 3 microsatélites) de la región flanqueante en un total de 687 personas no emparentadas de distintas poblaciones. Los SNPs fueron genotipados mediante unos ensayos de espectrometría de masa (tecnología MassARRAY® de

Sequenom), mientras que los microsatélites mediante un análisis de fragmentos usando la tecnología de Applied Biosystems.

Los análisis estadísticos incluyeron pruebas exactas para investigar la diferenciación poblacional y los patrones de desequilibrio de ligamiento en distintas zonas geográficas. La búsqueda de señales de selección positiva se llevó a cabo a través de dos pruebas diferentes: por un lado, se hizo una comparación de valores F_{ST} entre los marcadores neutros y de riesgo cardiovascular, y, por otro, se analizaron los patrones de homocigosidad haplotípica extendida (extended haplotype homozygosity – EHH).

Nuestros resultados confirman la insólita falta de desequilibrio de ligamiento entre las mutaciones adyacentes -402 y -402. Respecto a las pruebas de selección, no se ha detectado ninguna señal de selección positiva en la zona del mediterráneo o de África subsahariana. En cambio, algunos de los datos sugieren una potencial señal de selección positiva en los amerindios de Bolivia.

En conclusión, nuestros datos sugieren, aunque no demuestran, que posiblemente ha habido diferentes historias evolutivas en la región promotora del gen F7 en poblaciones mediterráneas y amerindias.

Supervisor's report on the involvement of the PhD student in the development of this paper

Dr. **Pedro Moral Castrillo**, Professor at the Department of Animal Biology of the University of Barcelona and supervisor of the doctoral thesis "Genetic variation of the X chromosome and the genomic regions of Coagulation Factors VII and XII in human populations: Epidemiological and evolutionary considerations" by **Georgios Athanasiadis**, hereby certifies that the participation of the above student in the article "**Different evolutionary histories of the coagulation factor VII gene in human populations?**", published in *Annals of Human Genetics,* consisted of the following tasks:

- DNA extraction from the Basque Country sample and template DNA preparation, in collaboration with Magda Gayà-Vidal

- Participation in the design of the study and selection of the analysed markers together with Dr. Pedro Moral Castrillo

- Genotype determination of the microsatellite polymorphisms in the lab

- Creation of the genotype database and statistical analysis of the data

- Participation in the manuscript drafting together with Dr. Pedro Moral Castrillo and Dr. Mark Stoneking

In addition, it is important to note that none of the co-authors of this article have used the results of this work in any implicit or explicit way to develop another doctoral thesis. As a consequence, this article forms part of the doctoral thesis of Georgios Athanasiadis exclusively.

Signed by Dr. Pedro Moral Castrillo

Barcelona, 9 April 2010

Annals of human genetics

doi: 10.1111/j.1469-1809.2009.00557.x

Different Evolutionary Histories of the Coagulation Factor VII Gene in Human Populations?

Georgios Athanasiadis[1], Esther Esteban[1], Magdalena Gayà-Vidal[1], Jean-Michel Dugoujon[2], Nicholas Moschonas[3], Hassen Chaabani[4], Nisrine Bissar-Tadmouri[5], Nourdin Harich[6], Mark Stoneking[7] and Pedro Moral[1]*

[1]*Unitat d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain*
[2]*Laboratory AMIS, University Toulouse III, Toulouse, France*
[3]*Laboratory of General Biology, School of Medicine, University of Patras, Patras, Greece*
[4]*Faculté de Pharmacie de Monastir, Université de Monastir, Monastir, Tunisia*
[5]*Sharjah University, College of Medicine, United Arab Emirates*
[6]*Deparament de Biologie, Université Chouaïb Doukkali, Morocco*
[7]*Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany*

Summary

Immoderate blood clotting constitutes a risk factor for cardiovascular disease in modern industrialised societies, but is believed to have conferred a survival advantage, i.e. faster recovery from bleeding, on our ancestors. Here, we investigate the evolutionary history of the Coagulation Factor VII gene (F7) by analysing five cardiovascular-risk-associated mutations from the F7 promoter and nine neutral polymorphisms (six SNPs and three microsatellites) from the flanking region in 16 populations from the broader Mediterranean region, South Saharan Africa and Bolivia (687 individuals in total). Population differentiation and selection tests were performed and linkage disequilibrium patterns were investigated. In all samples, no linkage disequilibrium between adjacent F7 promoter mutations −402 and −401 was observed. No selection signals were detected in any of the samples from the broader Mediterranean region and South Saharan Africa, while some of the data suggested a potential signal of positive selection for the F7 promoter in the Native American samples from Bolivia. In conclusion, our data suggest, although do not prove, different evolutionary histories in the F7 promoter region between Mediterraneans and Amerindians.

Keywords: Mediterranean region, coagulation factor VII, cardiovascular risk, human populations, long-range haplotype test

Introduction

Coagulation is a complex process by which blood forms clots, playing an important role in haemostasis. Disorders of coagulation can lead to an increased risk of bleeding or clotting. Mutations that lead to more efficient blood clotting are thought to have conferred a survival advantage on our ancestors and, thus, they must have been favoured by natural selection. Such an advantage could be a faster recovery from postpartum haemorrhage and posttraumatic bleeding. This hypothesis has been proposed for Factor V Leiden and could be true for variants in other blood clotting factors (Lindqvist & Dahlbäck, 2008). However, in modern industrialised societies this advantage seems to be overturned, as immoderate blood clotting constitutes a risk factor for cardiovascular disease (Mackay & Mensah, 2004).

One of the most thoroughly studied blood clotting elements is Coagulation Factor (F) VII, a vitamin K-dependent glycoprotein that is synthesised in the liver and secreted into the blood as an inactive zymogen (Fair, 1983). After endothelial damage, Tissue Factor is exposed and binds to FVII, initiating the Tissue Factor (extrinsic) pathway of the coagulation cascade. Elevated FVII plasma levels significantly increase the risk for cardiovascular disease (de Maat et al., 1997).

FVII plasma levels have high heritability (Souto et al., 2000), determined primarily by the F7 gene on chromosome

*Corresponding author: Prof. Pedro Moral, Ph.D., Dpt. Biologia Animal-Antropologia, Fac. Biologia, Universidad de Barcelona, Av. Diagonal 645, 08028 Barcelona (Spain). Tel: +34 934021461; Fax: +34 934035740; E-mail: pmoral@ub.edu

13 (Soria et al., 2005). In the recent years, many studies have been carried out on this gene in order to find the genetic determinants of variation in FVII plasma levels. Although many candidate polymorphisms were identified from across the entire gene, it is now believed that variability in FVII plasma levels is chiefly the result of regulatory non-coding and intronic variants, rather than amino acid changes (Sabater-Lleal et al., 2006).

From the identified polymorphisms located in the F7 promoter region, five are the object of the present study: base substitutions −670A>C, −630A>G, −402G>A, −401G>T and a 10bp insertion/deletion polymorphism at position −324 (in some studies cited at position −323). This last polymorphism is commonly referred to as a 10 bp "insertion", and for the sake of comparability with other studies we use the same term. Nonetheless, sequence comparisons between humans and other primates revealed that the insertion is actually the ancestral allele. In humans the derived 10 bp deletion is now the major allele with frequencies varying from 0.65, in a South Saharan population from Cameroon, to 0.99, in Papua New Guinea (Hahn et al., 2004). The high frequency of the 10 bp deletion suggests that this mutation arose early in human evolution.

Allele −670C and the 10 bp insertion, when acting separately, affect the phenotype by decreasing the FVII plasma levels (Sabater-Lleal et al., 2007). Conversely, alleles −630G, −402A and −401T lead to an increase of FVII plasma levels (Sabater-Lleal et al., 2007). However, the 10 bp insertion has a dominant effect over all others, lowering significantly the FVII plasma levels when present (Sabater-Lleal et al., 2007). Linkage disequilibrium (LD) precludes most of the above variants from manifesting their individual effect.

Some of these risk variants in the F7 promoter were identified through the comparison of individuals affected by multiple thrombotic events with unaffected individuals in a set of families from Spain (Souto et al., 2000). This kind of study, essential as it is to the discovery of genetic risk factors, does not provide any information about variation patterns in general populations; the study of the relative importance of selection and demography in general populations is essential to understanding the evolutionary history of risk loci. Cardiovascular disease usually has a late onset, allowing risk mutation carriers to reproduce. In such cases, factors like random genetic drift may matter more than selection. Thus, the resulting incidence and distribution of the disease will be influenced by population processes, which include structure and history (e.g. founder effects).

In this light, this work aims to describe the evolutionary history of the F7 region in human populations from around the Mediterranean. The Mediterranean region exhibits a remarkably heterogeneous pattern of cardiovascular incidence, with low mortality rates in western Europe but much higher mortality rates in both eastern Europe and North Africa (WHO Statistical Information System – http://www.who.int/whosis/en/, 2002). An interesting question is whether this heterogeneity in cardiovascular disease incidence is also reflected in the genomic structure of the F7 region in the Mediterranean populations.

Most previous studies of F7 variation focused on the coding sequence and did not include flanking variation (Souto et al., 2000; Soria et al., 2005; Sabater-Lleal et al., 2006; Sabater-Lleal et al., 2007). In a previous study, a comparison of risk variation from the F7 gene with neutral variation from different chromosomes was carried out in six Old World populations (Hahn et al., 2004). In the present study, however, we used neutral variation from the same genomic region of the F7 gene, in order to minimise any potential heterogeneity caused by different chromosomal regions. Our goal was to analyse the distribution of the five mutations mentioned above in the context of the surrounding neutral variation in order to investigate the evolutionary history of F7 variation in human populations from the broader Mediterranean region and other parts of the world.

Materials and Methods

Samples

A set of 16 human populations from different locations were analysed, thirteen of them originating from seven countries surrounding the Mediterranean Sea: Spain, France, Greece, Turkey, Morocco, Algeria and Tunisia. In Spain, samples were collected from the north (Asturias, Basque Country, Pas Valley), northeast (Catalonia) and south of the country (Andalusia). France was represented by a southern sample from Toulouse, while Greece by a sample from Crete and Turkey by an urban population from Istanbul. As for North Africa, the samples included 3 Berber ethnic groups from Morocco (Asni and Khenifra from High Atlas; Bouhria from Nostheast Atlas), one Berber group from Algeria (M'zab) and an urban group from Tunisia (Monastir). To provide a broader context, three non-Mediterranean groups (South Saharans from the Ivory Coast; Aymaras and Quechuas from Bolivia) were included in the analysis. Sample sizes ranged from 41 to 45 individuals, with the exception of the samples from Turkey and Algeria (n = 34 and 31 respectively), adding up to a total of 687 individuals. Blood samples were collected for DNA extraction from healthy and unrelated individuals of both sexes and all participants had their four grandparents born in the same region. A detailed geographic location of the samples from around the Mediterranean is shown in Figure 1. The study was performed in accordance with the guidelines of the Ethical Committee of the University of Barcelona and with informed consent of all the participants.

Figure 1 Geographic location of the populations studied around the Mediterranean. Sample size (in brackets) and abbreviations: AN: Asni, High Atlas, Morocco (44), AS: Oviedo, Asturias, Spain (45), BA: Basque Country, Spain (45), BO: Bouhria, Northeast Atlas, Morocco (45), CR: Crete Island, Greece (45), CT: Catalonia, Spain (45), KH: Khenifra, High Atlas, Morocco (44), MZ: M'zab, Algeria (41), PA: Pas Valley, Cantabria, Spain (44), SF: Toulouse, South France (45), SS: South Spain, Andalusia (45), TN: Monastir, Tunisia (42), TU: Istanbul, Turkey (34).

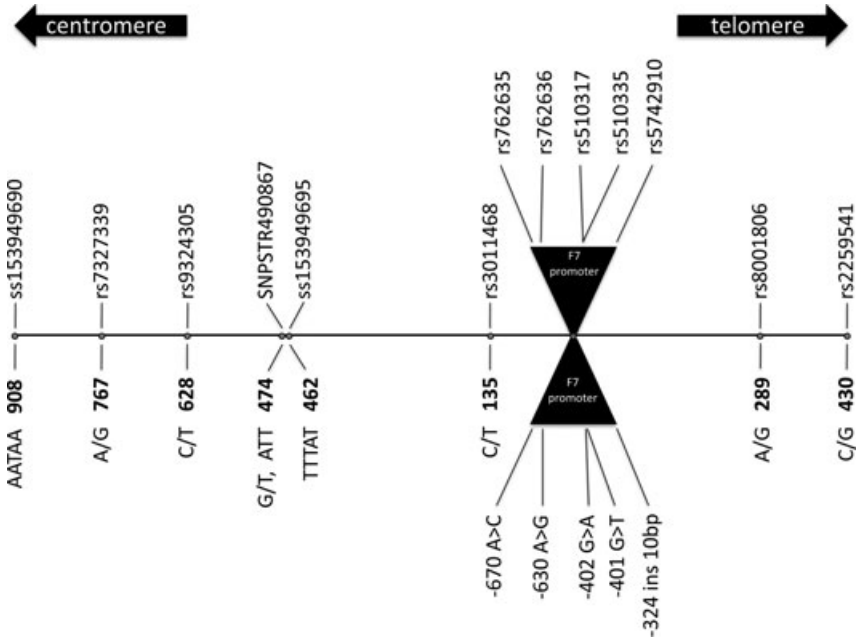


Figure 2 Schematic map of the neutral and risk markers in the F7 genomic region. The three-figure numbers in bold indicate distance from the gene (in Kbp).

Polymorphisms

Different kinds of genetic polymorphisms, characterised by distinct mutation mechanisms and rates, were considered in order to gain a comprehensive insight into the evolutionary history of the F7 region. The risk variants from the F7 promoter region included four SNPs and one insertion/deletion polymorphism. Five more SNPs were chosen from the wider genomic region of the F7 gene, up to a maximum distance of 1 Mb in both directions. These SNPs were located outside of any known genes or regulatory regions and thus were considered neutral. Moreover, these polymorphisms were selected according to the criterion of high heterozygosity in the CEU population (US residents of northern and western European ancestry) as reported in the HapMap project (www.hapmap.org). The reference ID of all SNPs analysed, as well as their relative position in the F7 genomic region, is shown in Figure 2.

In addition, three novel microsatellite loci from the same genomic region were analysed (Fig. 2). These markers were selected as follows: the F7 genomic region was scanned for microsatellites (tri-, tetra- and pentanucleotides) up to a maximum distance of 1 Mb in both directions from the gene, using the Tandem Repeats Finder algorithm (Benson, 1999) from the UCSC Genome Browser website (http://genome.ucsc.edu/cgi-bin/hgGateway). As the search brought out many results, the microsatellites were filtered on the basis of sequence purity (only perfect microsatellites were considered). The results from the analysis of the three microsatellites were submitted to GenBank and will become available in dbSNP Build 131.

Apart from SNPs, insertion/deletion polymorphisms and microsatellites, a SNPSTR was also considered in this study. SNPSTRs are a relatively new type of compound marker, consisting of one microsatellite and one tightly-linked SNP. As a consequence, SNPSTRs act like small haplotypes–a promising application in human population genetics (Mountain et al., 2002). For the identification of such loci in the F7 genomic region, the three selected microsatellites were scanned for nearby SNPs using the SNPSTR database (Agrafioti & Stumpf, 2007). The search returned SNP rs7994991, only 63 bp away from microsatellite $(AAT)_n$; the two markers together form SNPSTR490867.

Genotype Determinations

All SNPs and the insertion/deletion polymorphism were typed with the iPLEX™ Gold assay on the MassARRAY® platform (Sequenom, San Diego, CA, USA). For the microsatellites, PCR amplification was carried out using 1 μl of template (10–20 ng/μl), 1 μl Buffer 10x, 0.8 μl $MgCl_2$ (25 mM), 0.2 μl dNTPs (10 mM), 0.1 μl primers (25 μM) and 0.05 μl AmpliTaq Gold™ DNA Polymerase (Applied Biosystems, Foster City, CA, USA) in a total volume of 10 μl. The amplification program consisted of 2 min at 92°C, followed by 30 cycles of 20 s at 92°C, 20 s at the annealing temperature (varying from 58 to 66°C) and 20 s at 72°C, and with a final 5 min extension time at 72°C. One member of each primer pair was labelled at the 5' end with either the 6-FAM or HEX fluorescent dye. PCR was followed by 1:5 standard dilution and fragment analysis with the Applied Biosystems 3130 Genetic Analyzer using GeneScan™ 500 ROX™ as size standard. Genotypes were determined using the ABI Prism™ GeneMapper® v3.0 software (Applied Biosystems).

Statistical Analysis

Allele frequencies were calculated with GENETIX v4.05.2 (Belkhir et al., 1998) and genotype frequencies were tested for goodness-of-fit to Hardy-Weinberg proportions with ARLEQUIN v3.1 (Excoffier et al., 2005). Microsatellite statistics (number of alleles, mean, variance and heterozygosity) were calculated with Microsatellite analyzer (MSA) v4.05 (Dieringer & Schlötterer, 2003). Haplotype phase was inferred using PHASE v2.1 (Stephens et al., 2001; Stephens & Donnelly, 2003). Log ratio (G) tests were performed to detect significant population differences for each locus using GENEPOP v4.0 (Raymond & Rousset, 1995). Pairwise LD was assessed by D' and r^2 measures with Haploview v4.1 (Barrett et al., 2005) and by the Black and Krafsur test (Black & Krafsur, 1985) with GENETIX. The statistical significance of the non-random distribution of each pair of loci was tested by Fisher's Exact test with GENEPOP. The Bonferroni correction was applied in all cases of multiple testing.

Potential selective effects were investigated by F_{ST} and haplotype homozygosity analyses. F_{ST} values (Wright, 1951) were calculated for all loci by an analysis of molecular variance (AMOVA) using ARLEQUIN. The F_{ST} values from the risk polymorphisms were compared with those from the neutral loci by the non-parametric Mann-Whitney test (Mann & Whitney, 1947) in R (http://www.r-project.org). In the absence of selection, F_{ST} values reflect divergence in allele frequencies due to genetic drift.

Haplotype homozygosity was investigated via a long-range haplotype (LRH) test (Sabeti et al., 2002), which was carried out using the web-based tool found at http://ihg2.helmholtz-muenchen.de/cgi-bin/mueller/webehh.pl. The method involves identifying haplotypes in a locus of interest (core haplotypes) and following the decay of their association with other alleles at various distances from the locus. This association is evaluated by the extended haplotype homozygosity (EHH), an effective measure of LD for more than 2 markers (Sabeti et al., 2002). Core haplotypes with unusually high EHH and a high population frequency indicate positive selection. In our study, the core region consisted of the five risk variants and the analysis was carried out for the pooled samples from South Europe, North Africa, South Saharan Africa and Bolivia. It is important to note that in this method only SNP data were used.

Results

Allele Frequencies

After Bonferroni correction, none of the markers showed a significant departure from Hardy-Weinberg equilibrium in any population (data not shown). Allele frequencies of the risk and neutral biallelic polymorphisms from the F7 genomic region are shown in Table 1. Neutral SNPs rs799499, rs800180 and rs225954 present notable frequency differences across the Mediterranean, South Saharan and Amerindian samples. Polymorphisms −324, −401 and −402 from the Ivory Coast and the Mediterranean countries present similar frequencies to those previously found in Cameroon and Italy respectively (Hahn et al., 2004). The risk variants showed no clear differentiation pattern.

Allele frequencies of the three novel microsatellite polymorphisms are shown in Table 2 and a summary of their most important variation statistics is shown in Table 3. The variability of the three microsatellites is low to intermediate. Microsatellite $(AATAA)_n$ shows the lowest variability (five alleles, total $H = 0.300$), while $(AAT)_n$ is somewhat more

Table 1 Population allele frequencies (second row for each SNP) and heterozygosities (third row and in italics) of the 10 SNPs and one insertion/deletion polymorphism.

	N Spain	NE Spain	Pas Valley	S Spain	Basque Country	S France	Crete	Turkey	Asni Mor	Bouhria Mor	Khenifra Mor	M'zab Alg	Tunisia	Aymara	Quechua	Ivory Coast
rs732733	(44)	(43)	(43)	(45)	(43)	(44)	(45)	(32)	(43)	(43)	(43)	(29)	(39)	(42)	(44)	(41)
A/G	0.341	0.337	0.372	0.344	0.333	0.284	0.356	0.281	0.209	0.221	0.198	0.224	0.316	0.119	0.205	0.244
	0.455	*0.452*	*0.473*	*0.457*	*0.450*	*0.411*	*0.463*	*0.411*	*0.335*	*0.348*	*0.321*	*0.354*	*0.438*	*0.212*	*0.329*	*0.373*
rs932430	(44)	(43)	(43)	(45)	(43)	(44)	(45)	(32)	(43)	(43)	(43)	(29)	(39)	(42)	(43)	(41)
C/T	0.364	0.395	0.407	0.422	0.369	0.330	0.478	0.328	0.407	0.372	0.372	0.414	0.421	0.143	0.151	0.39
	0.468	*0.484*	*0.488*	*0.493*	*0.471*	*0.447*	*0.505*	*0.448*	*0.488*	*0.473*	*0.473*	*0.494*	*0.494*	*0.248*	*0.260*	*0.482*
rs799499	(42)	(45)	(43)	(45)	(44)	(41)	(43)	(32)	(44)	(42)	(43)	(30)	(41)	(45)	(44)	(44)
G/T	0.036	0.044	0.035	0.022	0.012	0.049	0.023	0.078	0.091	0.060	0.093	0.100	0.038	0.000	0.000	0.432
	0.070	*0.086*	*0.068*	*0.044*	*0.023*	*0.094*	*0.046*	*0.146*	*0.167*	*0.113*	*0.171*	*0.183*	*0.073*	*0.000*	*0.000*	*0.496*
rs301146	(44)	(43)	(43)	(45)	(43)	(44)	(45)	(32)	(43)	(43)	(43)	(29)	(39)	(42)	(43)	(42)
C/T	0.364	0.372	0.395	0.411	0.393	0.330	0.433	0.328	0.535	0.302	0.465	0.241	0.382	0.381	0.198	0.512
	0.468	*0.473*	*0.484*	*0.490*	*0.483*	*0.447*	*0.497*	*0.448*	*0.503*	*0.427*	*0.503*	*0.373*	*0.478*	*0.477*	*0.321*	*0.506*
rs762635	(44)	(45)	(42)	(45)	(43)	(43)	(44)	(31)	(41)	(42)	(42)	(27)	(40)	(41)	(43)	(39)
A/C	0.091	0.200	0.202	0.122	0.214	0.151	0.125	0.161	0.134	0.060	0.119	0.148	0.138	0.573	0.523	0.064
	0.167	*0.324*	*0.327*	*0.217*	*0.341*	*0.26*	*0.221*	*0.275*	*0.235*	*0.113*	*0.212*	*0.257*	*0.240*	*0.495*	*0.505*	*0.122*
rs762636	(42)	(41)	(43)	(45)	(41)	(41)	(43)	(32)	(44)	(42)	(43)	(29)	(41)	(45)	(44)	(44)
A/G	0.083	0.171	0.198	0.122	0.225	0.159	0.140	0.156	0.125	0.060	0.116	0.121	0.150	0.556	0.523	0.057
	0.155	*0.287*	*0.321*	*0.217*	*0.353*	*0.270*	*0.243*	*0.268*	*0.221*	*0.113*	*0.208*	*0.216*	*0.258*	*0.499*	*0.505*	*0.108*
rs510317	(45)	(45)	(43)	(45)	(43)	(45)	(42)	(33)	(42)	(43)	(42)	(28)	(42)	(43)	(42)	(44)
G/A	0.100	0.211	0.198	0.122	0.214	0.144	0.143	0.152	0.179	0.058	0.131	0.107	0.171	0.547	0.524	0.080
	0.182	*0.337*	*0.321*	*0.217*	*0.341*	*0.250*	*0.248*	*0.261*	*0.297*	*0.111*	*0.230*	*0.195*	*0.287*	*0.502*	*0.505*	*0.148*
rs510335	(45)	(45)	(43)	(45)	(42)	(44)	(45)	(31)	(42)	(43)	(44)	(29)	(40)	(39)	(39)	(42)
G/T	0.211	0.100	0.163	0.200	0.171	0.159	0.267	0.323	0.202	0.198	0.114	0.259	0.200	0.077	0.141	0.429
	0.337	*0.182*	*0.276*	*0.324*	*0.287*	*0.271*	*0.396*	*0.444*	*0.327*	*0.321*	*0.204*	*0.390*	*0.324*	*0.144*	*0.245*	*0.496*
rs574291	(45)	(45)	(43)	(45)	(43)	(44)	(45)	(31)	(42)	(43)	(44)	(28)	(41)	(39)	(40)	(41)
-/ins10bp	0.211	0.100	0.163	0.200	0.179	0.159	0.267	0.307	0.202	0.198	0.125	0.250	0.195	0.077	0.138	0.415
	0.337	*0.182*	*0.276*	*0.324*	*0.297*	*0.271*	*0.396*	*0.432*	*0.327*	*0.321*	*0.221*	*0.382*	*0.318*	*0.144*	*0.240*	*0.491*
rs800180	(42)	(45)	(43)	(45)	(44)	(41)	(43)	(32)	(44)	(42)	(43)	(30)	(41)	(45)	(43)	(45)
A/G	0.321	0.389	0.349	0.322	0.395	0.317	0.361	0.313	0.205	0.226	0.221	0.183	0.150	0.800	0.767	0.100
	0.442	*0.481*	*0.460*	*0.442*	*0.484*	*0.438*	*0.467*	*0.437*	*0.329*	*0.354*	*0.348*	*0.305*	*0.258*	*0.324*	*0.361*	*0.182*
rs225954	(45)	(45)	(43)	(45)	(43)	(44)	(44)	(31)	(42)	(43)	(44)	(28)	(40)	(38)	(38)	(42)
C/G	0.289	0.244	0.302	0.278	0.393	0.307	0.296	0.355	0.429	0.430	0.421	0.464	0.513	0.829	0.737	0.714
	0.416	*0.374*	*0.427*	*0.406*	*0.483*	*0.430*	*0.421*	*0.465*	*0.496*	*0.496*	*0.493*	*0.507*	*0.506*	*0.287*	*0.393*	*0.413*

The first row for each SNP shows number of individuals typed (in brackets). Polymorphisms are listed in the same order they are located on the chromosome towards the telomere. The featured frequencies correspond to the allele in bold.

Table 2 Population allele frequencies of the three microsatellite loci and SNPSTR490867.

	N Spain	NE Spain	Pas Valley	S Spain	Basque Country	S France	Crete	Turkey	Asni Mor	Bouhria Mor	Khenifra Mor	M'zab Alg	Tunisia	Aymara	Quechua	Ivory Coast
ss153949690																
$(AATAA)_3$	0.133	0.116	0.024	0.111	0.175	0.140	0.148	0.132	0.216	0.105	0.227	0.222	0.207	0.216	0.143	0.070
$(AATAA)_5$	0.856	0.872	0.976	0.889	0.825	0.826	0.852	0.868	0.705	0.826	0.705	0.759	0.744	0.784	0.857	0.826
$(AATAA)_6$											0.011					0.012
$(AATAA)_8$	0.011					0.023			0.068	0.058	0.034		0.049			0.070
$(AATAA)_9$		0.012				0.012			0.011	0.012	0.023	0.019				0.023
ss153949693																
$(AAT)_7$		0.012							0.012	0.012	0.046					
$(AAT)_8$		0.012														
$(AAT)_9$	0.279	0.174	0.186	0.133	0.261	0.256	0.167	0.188	0.186	0.171	0.193	0.179	0.075	0.128	0.036	0.390
$(AAT)_{10}$	0.074	0.012	0.014	0.022	0.022	0.033	0.064	0.063	0.128	0.110	0.125	0.036	0.050	0.012	0.024	0.354
$(AAT)_{11}$	0.632	0.779	0.729	0.844	0.717	0.700	0.731	0.688	0.651	0.683	0.614	0.750	0.838	0.663	0.833	0.171
$(AAT)_{12}$	0.015	0.012	0.071			0.011	0.039	0.063	0.023	0.012	0.011	0.036	0.038	0.198	0.107	0.085
$(AAT)_{13}$										0.012	0.011					
ss153949695																
$(TTTAT)_7$	0.346	0.232	0.370	0.219	0.357	0.435	0.269	0.292	0.281	0.200	0.256	0.296	0.061			0.250
$(TTTAT)_8$	0.462	0.446	0.389	0.594	0.452	0.457	0.558	0.458	0.438	0.700	0.524	0.523	0.939	0.790	0.941	0.250
$(TTTAT)_9$	0.115	0.143	0.056	0.109		0.022	0.077	0.083	0.078	0.050	0.073	0.091		0.026		0.139
$(TTTAT)_{10}$	0.039	0.107	0.037	0.016	0.048		0.019	0.083	0.016							0.028
$(TTTAT)_{11}$																0.056
$(TTTAT)_{12}$	0.039	0.054	0.019	0.031	0.048	0.022			0.047		0.061	0.046		0.026		0.194
$(TTTAT)_{13}$		0.018	0.074	0.016	0.048	0.065	0.019	0.042	0.109	0.050	0.061	0.023		0.158		0.083
$(TTTAT)_{14}$			0.056	0.016	0.048		0.058		0.031		0.024	0.023			0.059	
$(TTTAT)_{15}$								0.042								
SNPSTR490867																
G-$(AAT)_7$		0.011							0.011	0.011	0.046					
G-$(AAT)_9$	0.033	0.033	0.034	0.022	0.011	0.044	0.022	0.074	0.046	0.033	0.046	0.081	0.024			0.289
G-$(AAT)_{10}$									0.011							0.067
G-$(AAT)_{11}$	0.011									0.011						0.011
G-$(AAT)_{12}$									0.023			0.016	0.012			0.056
T-$(AAT)_8$		0.011														
T-$(AAT)_9$	0.178	0.144	0.125	0.111	0.122	0.211	0.133	0.103	0.136	0.133	0.148	0.097	0.049	0.122	0.034	0.089
T-$(AAT)_{10}$	0.056	0.011	0.011	0.022	0.011	0.033	0.056	0.059	0.114	0.100	0.125	0.032	0.049	0.011	0.023	0.256
T-$(AAT)_{11}$	0.711	0.778	0.773	0.844	0.844	0.700	0.756	0.706	0.648	0.689	0.614	0.758	0.842	0.678	0.841	0.211
T-$(AAT)_{12}$	0.011	0.011	0.057		0.011	0.011	0.033	0.059	0.011	0.011	0.011	0.016	0.024	0.189	0.102	0.022
T-$(AAT)_{13}$										0.011	0.011					

Table 3 Variation statistics of the three novel microsatellite loci.

	ss153949690 – $(AATAA)_n$					ss153949693 – $(AAT)_n$					ss153949695 – $(TTTAT)_n$				
	Sample Size	**N of Alleles**	**Mean**	**Variance**	**H**	**Sample Size**	**N of Alleles**	**Mean**	**Variance**	**H**	**Sample Size**	**N of Alleles**	**Mean**	**Variance**	**H**
N Spain	45	5	4.77	0.59	0.253	34	4	10.38	0.84	0.524	13	7	8.00	1.28	0.677
NE Spain	43	2	4.81	0.62	0.228	43	3	10.57	0.84	0.366	28	7	8.43	1.96	0.725
Pas Valley	41	4	4.95	0.10	0.048	35	5	10.69	0.74	0.436	27	7	8.54	4.52	0.711
S Spain	45	3	4.78	0.40	0.200	45	4	10.71	0.48	0.272	32	5	8.22	1.82	0.595
Basque Country	40	2	4.65	0.58	0.292	23	4	10.46	0.79	0.426	21	4	8.45	4.11	0.675
S France	43	4	4.84	0.94	0.302	45	6	10.47	0.79	0.448	23	4	8.00	2.49	0.611
Crete	44	2	4.70	0.51	0.255	39	4	10.64	0.65	0.438	26	6	8.29	2.95	0.618
Turkey	34	5	4.74	0.47	0.233	32	6	10.63	0.75	0.492	12	6	8.46	3.74	0.717
Asni Mor	44	3	4.82	1.64	0.457	43	4	10.48	0.82	0.531	32	6	8.75	4.54	0.719
Bourhia Mor	43	3	5.01	1.14	0.308	41	6	10.54	0.84	0.498	20	6	8.10	1.53	0.477
Khenifra Mor	44	2	4.75	1.55	0.455	44	4	10.34	1.26	0.575	41	7	8.51	3.49	0.654
M'zab Alg	27	2	4.63	1.07	0.381	28	4	10.64	0.67	0.41	22	2	8.23	2.51	0.643
Tunisia	41	3	4.73	1.21	0.406	40	4	10.84	0.37	0.293	41	8	9.44	5.64	0.691
Aymara	44	4	4.57	0.68	0.342	43	4	10.93	0.72	0.511	19	5	8.92	3.64	0.360
Quechua	42	2	4.71	0.5	0.248	42	4	11.01	0.28	0.296	17	6	8.35	2.05	0.114
Ivory Coast	43	2	5.17	1.28	0.312	41	3	9.95	0.91	0.695	18	6	9.50	5.07	0.838
Total	663	5	4.79	0.84	0.300	618	7	10.58	0.79	0.486	392	9	8.54	3.46	0.655

Total heterozygosity (H) refers to the heterozygosity of the pooled sample.

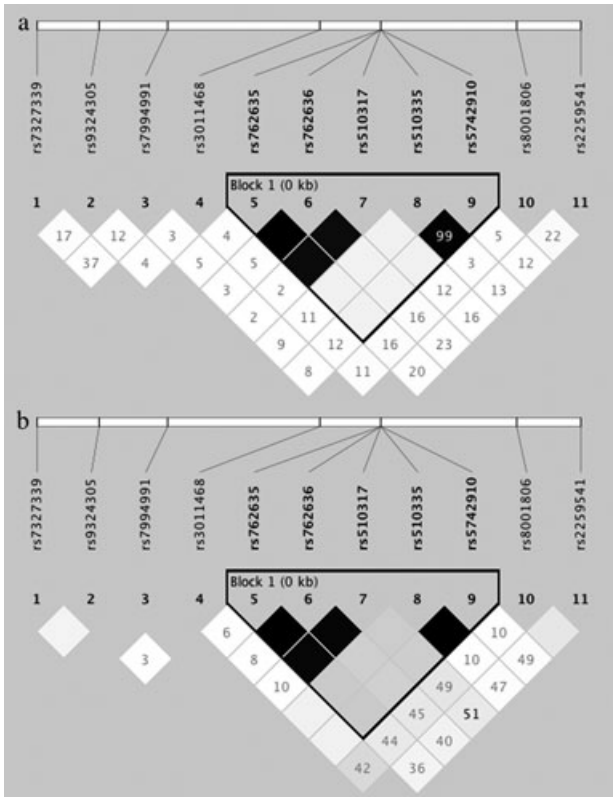


Figure 3 Haploview LD patterns for the biallelic markers studied in the F7 genomic region based on (a) all Mediterranean populations together and (b) the two Bolivian samples. Colour scheme represents r^2 values (white for $r^2 = 0$; shades of grey for $0 < r^2 < 1$; and black for $r^2 = 1$). Numbers represent D′ values (%). D′ in empty boxes equals 100% (D′ = 1).

variable (seven alleles, total H = 0.486). The highest variability is found in $(TTTAT)_n$ with 9 alleles and total H = 0.655.

Table 2 also shows the allele frequencies of SNPSTR490867. Here, allele T-$(AAT)_{11}$ predominates with frequencies ranging from 0.61 to 0.84 in all populations except the Ivory Coast, in which alleles G-$(AAT)_9$, T-$(AAT)_{10}$ and T-$(AAT)_{11}$ appear in almost equal frequencies.

Linkage Disequilibrium

In the pooled Mediterranean sample, the LD analysis revealed complete and statistically significant LD in all pairwise comparisons among the −670A>C, −630A>G and −402G>A polymorphisms (D′ = 1, $r^2 = 1$, Fisher's exact test, $p<0.001$) and also between −401G>T and −324ins10bp ((D′ = 0.99, $r^2 = 1$, Fisher's exact test, $p<0.001$; Fig. 3a). The same pattern was observed in the pooled Bolivian sample (Fig. 3b). In contrast, although D′ suggests complete LD between adjacent polymorphisms −402G>A and −401G>T, and in general between any marker from the first group with any marker from the second group, the r^2 measure indicates lack of linkage disequilibrium between them (e.g. for −402G>A and −401G>T: $r^2 = 0.04$ for the Mediterraneans and $r^2 = 0.15$ for the Bolivians). When Fisher's exact test was applied to these last cases, no statistically significant comparisons were observed. Both r^2 and Fisher's exact test suggest lack of LD between the adjacent polymorphisms −402G>A and −401G>T. Similar observations were reported in previous studies (Hahn et al., 2004; Soria et al., 2005; Sabater-Lleal et al., 2006). For the rest of the loci, no generalised LD patterns were observed.

Population Differentiation

When only the neutral variation was considered, the G tests showed that the three non-Mediterranean populations, from the Ivory Coast and Bolivia, were the most differentiated from all other populations. Focusing on the Mediterranean, no significant differences were found among South European samples for the neutral loci, while only two significant differences were observed among the North African samples for the same set of markers (those of Tunisia from Asni and Khenifra). The majority of the significant differences (11 out of 13) was observed between South European and North African samples. These population differentiation patterns are similar to previous results based on X chromosome *Alu* insertions (Athanasiadis et al., 2007) or autosomal *Alu* and *Alu*/microsatellite compound systems (González-Pérez et al., 2009). However, the differentiation patterns in the broader Mediterranean region, considering each risk variant separately, were not so clear. Approximately half of the significant pairwise population differences were found between South European and North African samples (6 out of 13 in −670; 7 out of 12 in −630; 4 out of 9 in −402; 7 out of 15 in −401; 5 out of 11 in −324ins10bp). The other half were found in almost equal parts within Europe and within Africa.

Global F_{ST} Comparisons

The global F_{ST} values of all risk and neutral markers are shown in Table 4. All F_{ST} values for the risk variants decreased when the Bolivian samples, first, and the South Saharans, second, were removed from the dataset. In the Mediterranean dataset, none of the F_{ST} values for the risk variants was significantly different from zero. On average, in none of the three datasets shown in Table 4 were risk F_{ST} values significantly greater than neutral F_{ST} values (data not shown). The observed lack of

	All populations	Old World	Mediterranean
$(AATAA)_n$	0.020	0.022	0.023
rs7327339	0.016	0.007*	0.007*
rs9324305	0.024	−0.004*	−0.004*
SNPSTR490867	0.069	0.069	0.011
$(TTTAT)_n$	0.055	0.038	0.025
rs3011468	0.022	0.017	0.013
−670A>C	**0.132**	**0.007***	**0.004***
−630A>G	**0.136**	**0.008***	**0.004***
−402G>A	**0.118**	**0.007***	**0.005***
−401G>T	**0.036**	**0.031**	**0.010***
−324ins10bp	**0.030**	**0.026**	**0.007***
rs8001806	0.160	0.030	0.019
rs2259541	0.094	0.048	0.015

Polymorphisms are listed according to their chromosomal order. An asterisk (*) indicates that the value is not significantly different from zero.

Table 4 Global F_{ST} values for the neutral and risk variants (the latter in bold), under three different datasets: All populations; Old World samples (i.e., excluding Bolivians); and Mediterranean samples only.

significant F_{ST} differences for the Mediterranean are in accord with a previous study that found no evidence for positive selection on polymorphisms −324, −401 and −402 in Italians (Hahn et al., 2004).

Analysis of Long-Range Haplotypes

Table 5 shows the frequencies of the core haplotypes in each population. Core haplotype (AAGG[-]) is predominant in all Old World populations. However, in the Native American samples, the most common haplotype is the one associated with higher cardiovascular risk (CGAG[-]). In South Europe and North Africa the higher and lower risk haplotypes present almost identical EHH patterns, as well as very similar frequencies (Fig. 4a). In the South Saharan sample, the lower risk haplotype shows a remarkably high frequency (0.40), but the EHH pattern is similar to that of the most common haplotype (Fig. 4b). Finally, in the Native American samples, there is a very different picture with the higher risk haplotype presenting both the highest frequency (0.54) as well as the highest EHH (Fig. 4c).

Discussion

This study explores the evolutionary history of the F7 genomic region in 16 worldwide populations, with a special focus on the Mediterranean basin, through the analysis of the molecular variation of five risk and eight neutral markers. Different types of genetic markers (SNPs, microsatellites and SNPSTRs) were selected to capture neutral variation. Our data dealt with some interesting aspects of the genetic architecture of the F7 genomic region and the processes that shaped diversity in it. To our knowledge, this is the most comprehensive general population study of the F7 genomic region in the broader Mediterranean region and the first one in Native American groups.

Genetic variation in the F7 promoter leads to variation in the FVII plasma levels (Souto et al., 2000). Since the latter were reported to affect susceptibility to cardiovascular disease (de Maat et al., 1997), the risk marker frequency variation found in our samples may be associated with different rates of cardiovascular incidence (http://www.who.int). However, our data are too few to provide definite conclusions, since the five risk loci correspond to only two inherited units (due to LD). Direct measurements of FVII plasma levels in these populations would be necessary to evaluate this hypothesis.

The lack of any signal of positive selection around the Mediterranean suggests that in this geographical region the variation on the F7 promoter did not bring any selective advantages to potential carriers. This gives support to the alternative view that the current F7 variation in the Mediterranean is the result of other diversity shaping factors like random genetic drift. In this light, these risk markers are essentially no different from neutral markers in this region. Indeed, in the Mediterranean populations, F_{ST} values for the risk markers did not differ significantly from the F_{ST} values for the neutral markers.

As the F_{ST} comparisons showed, the higher F_{ST} values of the −670C, −630G and −402A variants caused by the two Native American samples are better explained by genetic drift. However, the LRH test, which is more sensitive to recent selective events, revealed an unusual pattern for the risk haplotype in the Amerindians. It is important to note that this pattern cannot be interpreted as a solid proof of selection; there is a reasonable possibility for EHH/frequency

Table 5 F7 promoter core haplotype frequencies in 16 populations.

		Populations															
		N Spain	NE Spain	Pas Valley	S Spain	Basque Country	S France	Crete	Turkey	Asni Mor	Bouhria Mor	Khenifra Mor	M'zab Alg	Tunisia	Aymara	Quechua	Ivory Coast
Haplotypes	AAAG[-]		0.011							0.045		0.011		0.024		0.011	0.033
	AAAT[ins]							0.010							0.010		
	AAGG[ins]											0.011					
	AAGG[-]	0.689	0.689	0.648	0.678	0.600	0.700	0.590	0.559	0.636	0.756	0.750	0.629	0.634	0.370	0.341	0.511
	AAGT[ins]*	0.211	0.100	0.159	0.200	0.170	0.160	0.270	0.279	0.193	0.189	0.114	0.242	0.195	0.070	0.125	0.400
	AAGT[-]								0.015								
	AGGG[ins]						0.010										
	CAAT[ins]					0.020											
	CGAG[-]†	0.100	0.200	0.193	0.122	0.210	0.130	0.130	0.147	0.125	0.056	0.114	0.129	0.146	0.560	0.523	0.056

* Lower cardiovascular risk haplotype, †higher cardiovascular risk haplotype

differences in Amerindians to be attributed to drift, in full agreement with the F_{ST} findings mentioned above. In any case, our data leave open the possibility of selection of the F7 risk variants likely related to a more efficient recovery after injury (Lindqvist & Dahlbäck, 2008). In this context, signals of positive selection were reported in a Chinese sample from Singapore by the F_{ST} method (Hahn et al., 2004) and also a Yakut sample from the Sakha Republic in Siberia using the LRH test (G. Athanasiadis, pers. comm.).

Blood clotting is a complex trait. As such, many genetic factors affect the final phenotype. It follows that an advantageous mutation in a certain gene, unless affecting fitness in a dramatic way, will have few chances of being positively selected. In this context, we propose an explanation for our data. Transition G>A at locus −402 leads to higher FVII plasma levels and to more efficient blood clotting. However, in the Mediterranean region, this increased efficiency was either indifferent or was cancelled out by other unknown factors, as the lack of positive selection indicates. On the other hand, evidence of a signal of positive selection in some populations of Asian descent indicates that the same polymorphism under different circumstances might have had a powerful effect on fitness.

In conclusion, the potential protection from cardiovascular diseases that some Mediterranean groups enjoy (Tunstall-Pedoe, 2003) does not seem to be associated with any signature of positive selection involving the F7 promoter region. It may be that other loci are more important in providing protection against cardiovascular disease in these populations, or it may be that this reflects other processes than selection, such as genetic drift. Moreover, it is possible that the F7 promoter region presents different evolutionary histories between populations of Mediterranean and Asian descent, with potential consequences for cardiovascular susceptibility. To further investigate this observation, data from more loci and the analysis of cardiovascular incidence in more worldwide populations is desirable.

Acknowledgements

This research has been financially supported by the CGL2005-03391 and CGL2008-03955 projects of the Spanish Ministerio de Educación y Ciencia, the 2005SGR00252 project of the Generalitat de Catalunya and the HF2006-0210 grant. The work of GA has been financed by an FPU grant from the Ministerio de Educación y Ciencia (grant reference: AP2005-4425). We are grateful to all of the donors for providing blood samples and our collaborators from Spain (F Luna, C Rodríguez and M De Grado) for contributing to their collection.

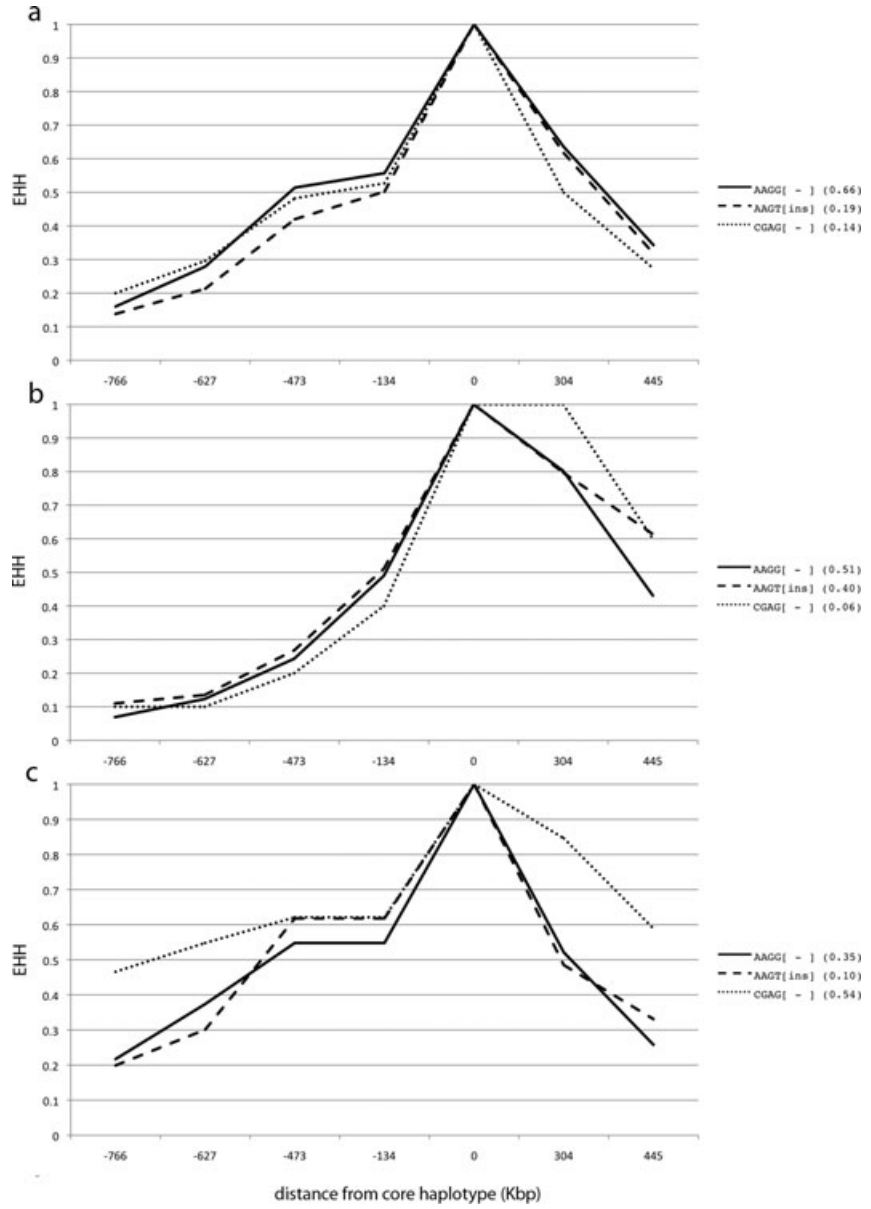


Figure 4 The EHH at varying distances from the core region, for each of the three most frequent core haplotypes for (a) the Mediterranean, (b) the Ivory Coast and (c) Bolivia. The value in brackets indicates haplotype frequency.

References

Agrafioti, I. & Stumpf, M. P. (2007) SNPSTR: a database of compound microsatellite-SNP markers. *Nucleic Acids Res* **35** (Database issue), D71–75.

Athanasiadis, G., Esteban, E., Via, M., Dugoujon, J. M., Moschonas, N., Chaabani, H. & Moral, P. (2007) The X chromosome Alu insertions as a tool for human population genetics: data from European and African human groups. *Eur J Hum Genet* **15**, 578–583.

Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265.

Belkhir, K., Borsa, P., Goudet, J., Chikhi, L. & Bonhomme, F. (1998) GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000. Université de Montpellier II, Montpellier, (France).

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580.

Black, W. C. & Krafsur, E. S. (1985) A FORTRAN program for the calculation and analysis of two-locus linkage disequilibrium coefficients. *Theor Appl Genet* **70**, 491–496.

de Maat, M. P., Green, F., de Knijff, P., Jespersen, J. & Kluft, C. (1997) Factor VII polymorphisms in populations with different risks of cardiovascular disease. *Arterioscler Thromb Vasc Biol* **17**, 1918–1923.

Dieringer, D. & Schlötterer, C. (2003) Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol Ecol Notes* **3**, 167–169.

Excoffier, L., Laval, G. & Schneider, S. (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinf Online* **1**, 47–50.

Fair, D. S. (1983) Quantification of factor VII in the plasma of normal and warfarin-treated individuals by radioimmunoassay. *Blood* **62**, 784–791.

González-Pérez, E., Esteban, E., Via, M., Gaya-Vidal, M., Athanasiadis, G., Dugoujon, J. M., Luna, F., Mesa, M. S., Fuster, V., Kandil, M., Harich, N., Bissar-Tadmouri, N., Saetta, A. & Moral, P. (2009) Population relationships in the Mediterranean revealed by autosomal genetic data (Alu and Alu/STR compound systems). *Am J Phys Anthropol* (in press).

Hahn, M. W., Rockman, M. V., Soranzo, N., Goldstein, D. B. & Wray, G. A. (2004) Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. *Genetics* **167**, 867–77.

Lindqvist, P. G. & Dahlbäck, B. (2008) Carriership of Factor V Leiden and evolutionary selection advantage. *Curr Med Chem* **15**, 1541–1544.

Mackay, J. & Mensah, G. (eds.) (2004) The Atlas of Heart Disease and Stroke Geneva. Switzerland: (WHO).

Mann, H. B. & Whitney, D. R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* **18**, 50–60.

Mountain, J. L., Knight, A., Jobin, M., Gignoux, C., Miller, A., Lin, A. A. & Underhill, P. A. (2002) SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Res* **12**, 1766–1772.

Raymond, M. & Rousset, F. (1995) GENEPOP version 1.2 population genetics software for exact tests and ecumenicism. *J Hered* **86**, 248–249.

Sabater-Lleal, M., Chillón, M., Howard, T. E., Gil, E., Almasy, L., Blangero, J., Fontcuberta, J. & Soria, J. M. (2007) Functional analysis of the genetic variability in the F7 gene promoter. *Atherosclerosis* **195**, 262–268.

Sabater-Lleal, M., Almasy, L., Martínez-Marchán, E., Martínez-Sánchez, E., Souto, R., Blangero, J., Souto, J., Fontcuberta, J. & Soria, J. M. (2006) Genetic architecture of the F7 gene in a Spanish population: implication for mapping complex diseases and for functional assays. *Clin Genet* **69**, 420–428.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. & Lander, E. S. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837.

Soria, J. M., Almasy, L., Souto, J. C., Sabater-Lleal M., Fontcuberta J. & Blangero, J. (2005) The F7 gene and clotting factor VII levels: dissection of a human quantitative trait locus. *Hum Biol* **77**, 561–575.

Souto, J. C., Almasy, L., Borrell, M., Garí, M., Martínez, E., Mateo, J., Stone, W. H., Blangero, J. & Fontcuberta, J. (2000) Genetic determinants of hemostasis phenotypes in Spanish families. *Circulation* **101**, 1546–1551.

Stephens, M. & Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* **73**, 1162–1169.

Stephens, M., Smith, N. J. & Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978–989.

Tunstall-Pedoe, H. (ed.) Prepared by Tunstall-Pedoe H, Kuulasmaa K, Tolonen H, Davidson M, Mendis S with 64 other contributors for The WHO MONICA Project (2003) MONICA Monograph and Multimedia Sourcebook Geneva. Switzerland: (WHO).

Wright, S. (1951) The genetical structure of populations. *Ann Eugen* **15**, 323–354.

Received: 25 August 2009
Accepted: 10 October 2009

Results IV

Athanasiadis et al., 2010b

The Mediterranean Sea as a barrier to gene flow: evidence from variation in and around the F7 and F12 genomic regions

Georgios Athanasiadis, Emili Gonzalez-Perez, Esther Esteban, Jean-Michel Dugoujon, Mark Stoneking and Pedro Moral

BMC Evolutionary Biology 2010; 10(1): 84

Resumen en castellano

La región mediterránea se caracteriza por una larga historia de interacciones entre diferentes poblaciones. En este trabajo se estudian las relaciones genéticas entre 13 muestras de poblaciones procedentes de la región mediterránea, junto con otros grupos de la Costa de Marfil y Bolivia, prestando especial atención a la estructura genética entre el norte de África y el sur de Europa.

Las determinaciones analíticas incluyeron un diverso conjunto de polimorfismos neutros y funcionales situados dentro y alrededor de las regiones genómicas de los factores de coagulación VII y XII (genes F7 y F12). En total se analizaron 15 SNPs, 6 microsatélites y un polimorfismo de inserción/deleción (indel) de 10 pares de bases. Los SNPs y el indel fueron genotipados mediante unos ensayos de espectrometría de masa (tecnología MassARRAY® de Sequenom), mientras que los microsatélites mediante un análisis de fragmentos usando la tecnología de Applied Biosystems.

El análisis de componentes principales ha revelado una agrupación significativa de las muestras mediterráneas en grupos que geográficamente corresponden al

norte de África y el sur de Europa. Dicha agrupación es consistente con los resultados del análisis jerárquico de varianza molecular (hierachical analysis of molecular variance – AMOVA), el cual ha mostrado una baja pero significativa diferenciación entre los grupos de las dos costas mediterráneas. Un análisis más detallado de esta diferenciación, utilizando haplotipos, proporciona evidencias parciales de un flujo génico subsahariano más alto en el norte de África que en el sur de Europa.

Los resultados obtenidos de las dos regiones genómicas consideradas respecto al flujo génico a través del Sáhara no son coincidentes ya que mientras la región F12 indica que el norte de África comparte más haplotipos con la Costa de Marfil que el sur de Europa, en la región F7 no se han encontrado diferencias significativas entre los dos lados del Mediterráneo respecto a los haplotipos compartidos con poblaciones subsaharianas. Por tanto, es muy difícil llegar a una conclusión sólida sobre el papel del flujo génico en la diferenciación entre las dos costas mediterráneas, y mas datos son necesarios para llegar a una conclusión definitiva. Sin embargo, nuestros datos sugieren que el Mediterráneo ha sido al menos en parte una barrera al flujo génico entre las dos costas, que habría condicionado un menor flujo génico subsahariano a la costa europea del Mediterráneo.

Supervisor's report on the involvement of the PhD student in the development of this paper

Dr. **Pedro Moral Castrillo**, Professor at the Department of Animal Biology of the University of Barcelona and supervisor of the doctoral thesis "Genetic variation of the X chromosome and the genomic regions of Coagulation Factors VII and XII in human populations: Epidemiological and evolutionary considerations" by **Georgios Athanasiadis**, hereby certifies that the participation of the above student in the article "**The Mediterranean Sea as a barrier to gene flow: evidence from variation in and around the F7 and F12 genomic regions**", published in *BMC Evolutionary Biology,* consisted of the following tasks:

- DNA extraction from the Basque Country sample and template DNA preparation, in collaboration with Magda Gayà-Vidal

- Participation in the design of the study and selection of the analysed markers in together with Dr. Pedro Moral Castrillo

- Genotype determination of the microsatellite polymorphisms in the lab

- Creation of the genotype database and statistical analysis of the data

- Participation in the manuscript drafting together with Dr. Pedro Moral Castrillo and Dr. Mark Stoneking

In addition, it is important to note that none of the co-authors of this article have used the results of this work in any implicit or explicit way to develop another doctoral thesis. As a consequence, this article forms part of the doctoral thesis of Georgios Athanasiadis exclusively.

Signed by Dr. Pedro Moral Castrillo

Barcelona, 9 April 2010

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

The Mediterranean Sea as a barrier to gene flow: evidence from variation in and around the F7 and F12 genomic regions

Georgios Athanasiadis[1], Emili González-Pérez[1], Esther Esteban[1], Jean-Michel Dugoujon[2], Mark Stoneking[3], Pedro Moral[1*]

Abstract

Background: The Mediterranean has a long history of interactions among different peoples. In this study, we investigate the genetic relationships among thirteen population samples from the broader Mediterranean region together with three other groups from the Ivory Coast and Bolivia with a particular focus on the genetic structure between North Africa and South Europe. Analyses were carried out on a diverse set of neutral and functional polymorphisms located in and around the coagulation factor VII and XII genomic regions (F7 and F12).

Results: Principal component analysis revealed a significant clustering of the Mediterranean samples into North African and South European groups consistent with the results from the hierarchical AMOVA, which showed a low but significant differentiation between groups from the two shores. For the same range of geographic distances, populations from each side of the Mediterranean were found to differ genetically more than populations within the same side. To further investigate this differentiation, we carried out haplotype analyses, which provided partial evidence that sub-Saharan gene flow was higher towards North Africa than South Europe.

Conclusions: As there is no consensus between the two genomic regions regarding gene flow through the Sahara, it is hard to reach a solid conclusion about its role in the differentiation between the two Mediterranean shores and more data are necessary to reach a definite conclusion. However our data suggest that the Mediterranean Sea was at least partially a barrier to gene flow between the two shores.

Background

The history of the Mediterranean involves successive population movements across the lands that surround it, both in prehistoric and historical times. In historical times, these population movements have included peoples like Greeks, Romans, Celts, Goths, Slavs, Arabs and Turks[1]. It is thus a great challenge - as the great number of relevant human population genetic studies also reveals - to investigate the extent to which this intense migratory activity has influenced the genetic composition of the present Mediterranean populations.

Regarding the Mediterranean genetic profile, a recent X chromosome SNP study showed that the region exhibits a high overall genetic homogeneity,[2] which seems to agree with an apparently weak genetic structure between South Europeans and North Africans, as revealed by an analysis of Y chromosome microsatellites[3]. This pattern may be a consequence of the Neolithic demic diffusion in this region (around 10,000 years before present) and/or a high level of gene flow in the area.

In any case, the genetically homogeneous Mediterranean landscape is sprinkled with differentiated isolates such as the Corsicans,[4] the Sardinians[5] and populations from the Balearic Islands[6]. Moreover, a Moroccan sample was found to present significant genetic differences from other Mediterranean populations in their X chromosomes[2]. This last observation has been attributed by some scholars to the potential role of the Gibraltar Strait as a genetic barrier between Northwest Africa and the Iberian Peninsula, [7] although there is no general consensus on this issue, [8,9] possibly reflecting the fact that different markers and genomic components reveal different patterns.

* Correspondence: pmoral@ub.edu
[1]Unitat d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain

In this study we investigate the genetic structure of human populations in the Mediterranean, with a particular emphasis on the genetic relationships between groups from North Africa and South Europe. We paid special attention to the role of gene flow through the Sahara in the genetic differentiation between Northern Africans and Southern Europeans. To accomplish our goals, we used polymorphisms in and around the genomic regions of the F7 and F12 genes. These genes code for the coagulation factors VII and XII respectively and are involved in blood clotting. The chosen polymorphisms from the functional regions of the two genes were previously reported to be associated with susceptibility to cardiovascular disease in groups from the Mediterranean[10,11].

Some of the data used here (i.e. variation in and around the F7 gene) were published previously,[12] while new data include neutral variation around the F12 gene and the F12 46C>T functional polymorphism. This extensively studied marker is related to Factor XII plasma levels and the development of thrombosis, although the causal relationship between these two features is questionable[13].

According to our data, the Mediterranean populations are significantly clustered into South Europeans and North Africans, despite the low genetic differentiation between the two groups. Our analyses also suggest that this differentiation can be explained by the Mediterranean Sea acting a genetic barrier, which may also have affected the sub-Saharan gene flow into the Mediterranean region.

Methods

Samples

A set of 16 human populations (687 individuals) from different locations were analysed, thirteen of them originating in seven countries from around the Mediterranean: Spain (Asturias, Basque Country, Pas Valley in the north; Catalonia in the northeast; Andalusia in the south), France (Toulouse in the south), Greece (Crete island), Turkey (Istanbul), Morocco (Asni and Khenifra Berbers from High Atlas; Bouhria Berbers from Northeast Atlas), Algeria (M'zab Berbers) and Tunisia (Monastir). The location of the Mediterranean samples is shown in Figure 1. In addition, three non-Mediterranean groups (sub-Saharan Africans from the Ivory Coast; Aymaras and Quechuas from Bolivia) were included in the analysis. Sample sizes ranged from 41 to 45 individuals, with the exception of the samples from Turkey and Algeria (n = 34 and 31 respectively). Blood samples were collected for DNA extraction from healthy and unrelated individuals of both sexes and all participants had their four grandparents born in the same region. The study was performed in accordance with the guidelines of the Ethical Committee of the University of Barcelona and with informed consent of all the participants.

Polymorphisms

In the present study, functional variation is represented by 4 SNPs and one insertion/deletion polymorphism from the F7 promoter region[12] and the 46C>T polymorphism (rs1801020) from the 5'-untranslated region in F12 exon 1,[11] also referred to as 'risk' markers. The study also included 5 more SNPs, 3 microsatellites and one SNPSTR from the wider genomic region of the F7 gene,[12] as well as 4 SNPs and 3 microsatellites from the wider genomic region of the F12 gene (Figure 2). These last 16 polymorphisms were located outside of any known genes or regulatory regions and, thus, were considered to be neutral. SNPs were selected for genotyping according to the criterion of high heterozygosity in the CEU population (US residents of northern and western European ancestry) as reported in the HapMap project http://www.hapmap.org, while microsatellites were selected as described previously[12].

Genotype determinations

All SNPs and the insertion/deletion polymorphism were typed with the iPLEX™ Gold assay on the Sequenom MassARRAY® Platform. For the microsatellites, PCR amplification was carried out, followed by 1:5 standard dilution and fragment analysis with the Applied Biosystems 3130 Genetic Analyzer[12].

Statistical analysis

Allele frequencies of all polymorphisms were calculated with GENETIX v4.05.2[14]. Genotype frequencies were tested for goodness-of-fit to Hardy-Weinberg proportions with ARLEQUIN v3.1[15]. Additional microsatellite statistics (number of alleles; mean and variance of repeat number; and heterozygosity) were calculated using Microsatellite Analyzer (MSA) v4.05[16].

In most of the analyses that follow, our intention was to use as many of the chosen polymorphisms as possible from both the F7 and F12 genomic regions. To achieve this, we first tested all 'risk' markers for selective neutrality. Those 'risk' markers for which neutrality could not be rejected were lumped together with the neutral ones. Neutrality was tested through F_{ST} comparisons: F_{ST} values[17] were calculated for all loci by a locus-by-locus analysis of molecular variance (AMOVA) using ARLEQUIN. The molecular distance used was the number of pairwise differences. In the absence of selection, the F_{ST} values from the 'risk' polymorphisms are expected to have the same distribution as the F_{ST} values from the neutral loci. For the F7 genomic region, neutral and 'risk' F_{ST} values were compared by a Mann-Whitney test[18] in R http://www.r-project.org. For the F12 genomic region, the single 'risk' 46C>T F_{ST} value was compared with the 95% confidence interval from the corresponding neutral F_{ST} distribution. As a final step, an additional

Figure 1 Geographic location of the populations studied around the Mediterranean. Sample size (in brackets) and abbreviations: AN: Asni, High Atlas, Morocco (44), AS: Oviedo, Asturias, Spain (45), BA: Basque Country, Spain (45), BO: Bouhria, Northeast Atlas, Morocco (45), CR: Crete Island, Greece (45), CT: Catalonia, Spain (45), KH: Khenifra, High Atlas, Morocco (44), MZ: M'zab, Algeria (31), PA: Pas Valley, Cantabria, Spain (44), SF: Toulouse, South France (45), SS: South Spain, Andalusia (45), TN: Monastir, Tunisia (41), TU: Istanbul, Turkey (34).

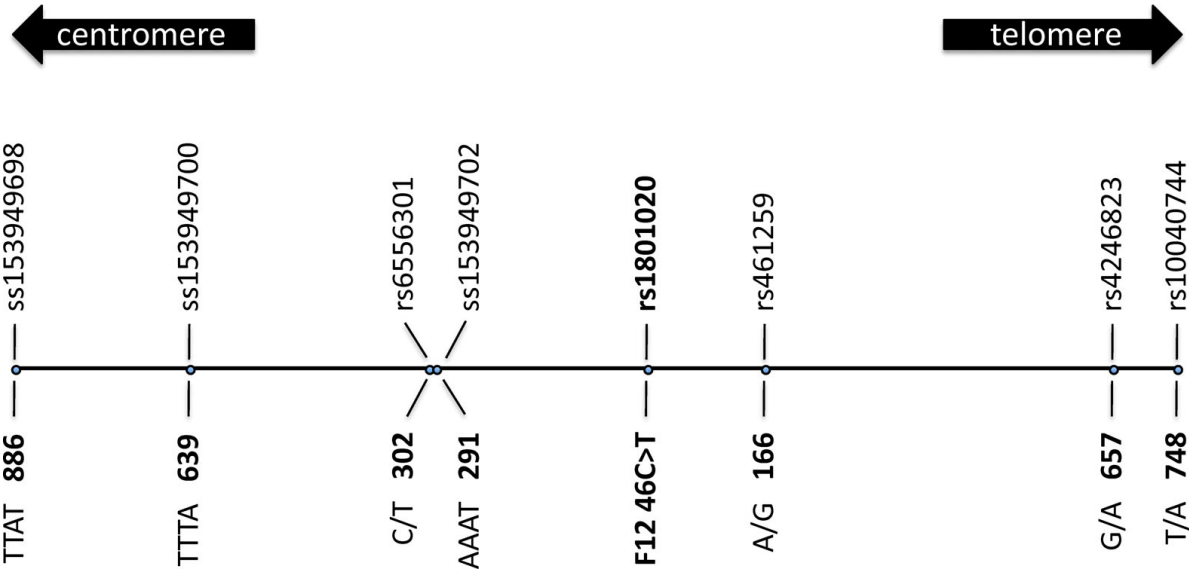


Figure 2 Schematic map of the neutral and risk markers in the F12 genomic region. The three-figure numbers in bold indicate distance from the gene (in Kbp).

Mann-Whitney test was carried out to check for significant differences in the patterns of variation between the markers from the two genomic regions.

In order to gain a first insight into the genetic relationships among our samples, we carried out a principal component analysis (PCA) with the *ade4* statistical package in R,[19] using the allele frequencies of all the markers from both the F7 and F12 regions. PCA was performed for 3 different datasets: (i) all 16 populations, (ii) Old World populations only (i.e. without the Bolivian samples) and (iii) populations from the broader Mediterranean region only. PC significance was evaluated through linear correlation of PC axes with group membership of each population by an analysis of variance (ANOVA) in R. PC eigenvalues were treated as dependent variables and group membership of each population as factors. The factors used were 'South Europe', 'North Africa', 'Ivory Coast' and 'South America'.

Population structure in our samples was also surveyed by a hierarchical AMOVA (molecular distance used: number of pairwise differences) using ARLEQUIN. The input files contained the genotypic data from both the F7 and F12 genomic regions. Isolation by distance (IBD) as a possible mechanism for the observed patterns of differentiation was evaluated by a Mantel test[20] of correlation between genetic and geographic distances using the *ade4* statistical package in R (10,000 permutations). The test was carried out for the 'Old World', 'Western Mediterranean', 'North Africa', 'South Europe' and 'Western Europe' sample subsets. The genetic distance used was that of Reynolds,[21] calculated from the allele frequencies of all the markers from both the F7 and F12 regions with PHYLIP v3.69[22]. Pairwise geographic distances (in Km) were calculated from the geographic coordinates (lat, lon) of each sample using the following formula:

$$\text{Distance} = 6378.137 \times \arccos[\sin(\text{lat}_1) \times \sin(\text{lat}_2) + \cos(\text{lat}_1) \times \cos(\text{lat}_2) \times \cos(\text{lon}_2\text{-lon}_1)]$$

To explore the possibility of a genetic boundary, we plotted together the geographic vs. genetic distances of both same-coast and opposite-coast pairs of Mediterranean samples. Genetic distances among same-coast samples similar to those among opposite-coast samples indicate that isolation by distance (IBD) is the most plausible model of differentiation. Conversely, greater genetic distances among opposite-coast samples as compared to same-coast samples would suggest that the Mediterranean is a barrier to gene flow.

We also searched for differences in gene flow from sub-Saharan Africa towards North Africa and South Europe as a potential consequence of the Mediterranean Sea acting as a genetic barrier via two different methods:

First, the degree of haplotype sharing among different groups was determined at a regional level (i.e. sub-Saharan Africa, North Africa, South Europe), as well as at a sample level (each sample treated individually); haplotypes based on both SNPs and microsatellites were inferred for each of the two genomic regions with PHASE v2.1[23,24] and were further analysed with ARLEQUIN to determine which haplotypes were shared by each pair of populations. This analysis was repeated for the SNPs alone. If the Mediterranean Sea was a genetic barrier between North Africa and South Europe, then we would expect a greater number of haplotypes to be shared between North Africans and sub-Saharans than between South Europeans and sub-Saharans, contributing also to a greater genetic differentiation between the two shores.

Second, pairwise Nm values,[25] which reflect the rate of migrant exchange between two populations, were estimated using the markers from both genomic regions with ARLEQUIN, applying the same molecular distance as in the AMOVA (see above). As ARLEQUIN returns the matrix of the *M values* (M = 2 Nm for diploid populations), we divided this matrix by a factor 2. Again, if the 'Mediterranean as a genetic barrier' hypothesis was true, we would expect higher Nm values between the Ivory Coast and each of the North African samples than between the Ivory Coast and the South Europeans. Conversely, similar numbers of shared haplotypes or Nm values would point towards the lack of a genetic barrier imposed by the Mediterranean Sea.

Results and Discussion

Allele frequencies, Hardy-Weinberg equilibrium and heterozygosity

After Bonferroni correction, none of the markers showed a significant departure from Hardy-Weinberg equilibrium in any population (data not shown). [Additional file 1] shows allele frequencies of the 'risk' and neutral SNPs from the F12 genomic region. The frequency of the 'risk' variant T in the polymorphism 46C>T (rs1801020) ranges from 0.081 to 0.357 in the samples from the Mediterranean countries, but was higher in the Ivory Coast (0.464) and the Native American samples (0.539 in Aymara and 0.577 in Quechua). The 46C>T frequency pattern in Native Americans was closer to that reported for Asians (C/T frequency = 0.27/0.73) than for Europeans[26]. Allele frequencies and variation statistics of the three novel microsatellites from the F12 genomic region are shown in [Additional file 2] and [Additional file 3]. Microsatellites $(TTAT)_n$ and $(AAAT)_n$ show moderate-high variability (9 alleles, total H ≈ 0.7), while tetranucleotide $(TTTA)_n$ is considerably less variable (6 alleles, total H = 0.181). The results from the analysis of the three microsatellites were submitted to GenBank and will become available in dbSNP Build 131. For the F7 genomic region, allele frequencies of the 'risk' and neutral biallelic, as well as microsatellite, polymorphisms were reported elsewhere[12].

F_{ST} comparisons

The F_{ST} values for the 'risk' and neutral markers from the F12 genomic region are shown in Table 1. As in the case of the F7 gene,[12] the F_{ST} value for the F12 46C>T 'risk' variant decreased when the Bolivian samples, first, and the Ivory Coast, second, were removed from the dataset. The 46C>T F_{ST} value was not significantly different from the neutral F_{ST} values in any of the three datasets considered (All Populations: 46C>T F_{ST} = 0.104, neutral F_{ST} 95% CI [-0.012, 0.116]; Old World: 46C>T F_{ST} = 0.047, neutral F_{ST} 95% CI [-0.015, 0.082]; Mediterranean: 46C>T F_{ST} = 0.020, neutral F_{ST} 95% CI [-0.015, 0.050]). This last observation suggests that polymorphism 46C>T can be treated as a 'neutral' polymorphism in all the analyses of population relationships. In the F7 gene, 'risk' and neutral F_{ST} values were also found not to differ significantly[12]. Finally, there seem to be no significant

Table 1 Global F_{ST} values for the neutral and risk variants (the latter in bold) of the F12 gene under 3 different datasets: All Populations; Old World samples (i.e., excluding Bolivians); and Mediterranean samples only

	All Populations	Old World	Mediterranean
$(TTAT)_n$	0.028	0.025	0.004*
$(TTTA)_n$	0.087	0.080	0.045
rs6556301	0.106	0.037	0.037
$(AAAT)_n$	0.036	0.023	0.013
46C>T	**0.104**	**0.047**	**0.020**
rs461259	0.030	0.008*	0.009*
rs4246823	0.022	0.012*	0.013*
rs10040744	0.053	0.048	0.002*

Polymorphisms are listed according to their chromosomal order. An asterisk (*) indicates that the value is not significantly different from zero.

differences in the patterns of variation between the two genomic regions in any of the three datasets (Mann-Whitney test, $p > 0.05$.), which allows the fusion of the two sets of markers in all the pertinent analyses.

Principal component analysis

Regarding population structure, when all of the populations were analysed, PCA identified three clusters, corresponding to: the Mediterranean groups; the two Bolivian groups; and the Ivory Coast (Figure 3a). In the Mediterranean cluster, all South European groups appeared separated from the North African groups along the first PC, except for Tunisia, which was closer to the South European groups. A similar result for the same Tunisian sample was reported in a previous study of Alu insertion polymorphisms on the X chromosome[27]. Although the first two PCs account for 39.52% of the original variation and the separation between North Africans and South Europeans along the second PC is visually not as clear, the population clustering according to the factors used ('South Europe', 'North Africa', 'Ivory Coast' and 'South America') was highly significant (Table 2). The ANOVA showed that the four geographical regions are clearly separated along the first PC. Populations are significantly separated along the second PC as well, although this separation is visually not as clear as in the first PC.

In the PCA plot of the old world dataset (Mediterranean countries and Ivory Coast), the 13 populations from around the Mediterranean are clustered together while the Ivory Coast is positioned further away (Figure 3b). This plot revealed a clear separation between the North Africans and South Europeans along both PC1 and PC2. The North African samples are slightly closer to the sub-Saharan sample than are the European samples in PC1, but not PC2. This may reflect a higher genetic affinity of sub-Saharan Africa with North Africa than with South Europe, due to a potentially higher gene flow from the south of the Sahara towards North Africa. Clustering by

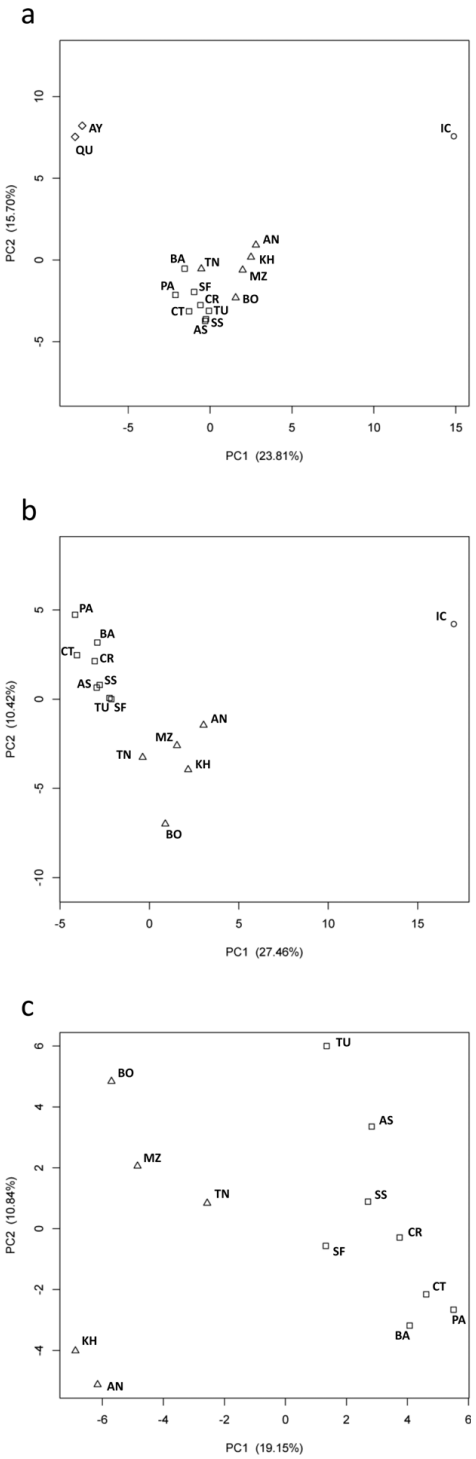


Figure 3 PCA plots of the first and second PCs for: (a) All samples; (b) the Old World samples (i.e. excluding Bolivians); (c) the Mediterranean samples only. The percentage of variation explained by each PC is shown in brackets. North African populations are represented by a triangle, South European populations by a square, Native Americans by a diamond and the Ivory Coast by a circle. PC Analyses were based on data from both the F7 and the F12 genomic regions.

Table 2 Analysis of variance for the PC significance

	All populations (S.E. - N.A. - I.C. - S.AM.)		Old world populations (S.E. - N.A. - I.C.)		Mediterranean populations (S.E. - N.A.)	
	F-quotient	p-value	F-quotient	p-value	F-quotient	p-value
PC 1	124.57	<0.01	195.68	<0.01	91.43	<0.01
PC 2	60.26	<0.01	16.22	<0.01	0.05	0.83
PC 3	15.98	<0.01	0.75	0.49	0.77	0.40
PC 4	0.09	0.97	0.35	0.71	<0.01	0.97

Abbreviations in the brackets refer to the geographical predictors used in each dataset. S.E.: South Europe, N.A.: North Africa, I.C.: Ivory Coast, S.AM.: South America.

population groups along the first two PCs (37.88% of the original variation) was again significant (Table 2).

In the PCA plot of just the Mediterranean dataset, the South Europe and North Africa clusters were maintained along the first PC (Figure 3c). Tunisia is the closest of the North African groups to the South Europe groups. The first two PCs account for 29.99% of the variance, although the population clustering into these two groups was significant only along the first PC (Table 2). It is also worth noting that the High Atlas Moroccans (Khenifra and Asni) seem to be separated from the other North Africans along the second PC.

Hierarchical AMOVA

Table 3 summarises the main findings from the hierarchical AMOVA. As the decreasing values of the F-statistics indicate, the Bolivians and Ivory Coast substantially contribute to the genetic variance found in our samples: the F_{ST} value decreases from 8.87% among all populations to 3.97% when the Bolivian groups are removed, to 1.67% when Ivory Coast is then removed. In the Mediterranean dataset, only 1.34% of the genetic variance was attributable to the South Europe vs. North Africa grouping (F_{CT} = 0.013, $p < 0.05$). Comas et al.[7] and González-Pérez et al.[8] had previously studied population relationships in the western Mediterranean using autosomal Alu polymorphisms, reporting a slightly higher differentiation (F_{CT} values: 0.020 and 0.018 respectively) as compared to our results. These findings suggest that the differentiation found between the two shores of the Mediterranean (also seen in all the above PC plots) is low, albeit significant.

Table 3 Hierarchical analysis of molecular variance based on the variation of both the F7 and the F12 genomic regions for three population datasets: All populations; Old World samples (i.e. excluding Bolivians); and Mediterranean samples only

	All	Old World	Mediterranean
Within populations	91.13	96.03	98.33
Among populations within groups	0.39	0.34	0.34
Among groups	8.52	3.64	1.34

The values correspond to the percentage of the variation explained by the corresponding grouping method. All values were statistically significant ($p < 0.05$).

Genetic-geographic correlation

Table 4 presents the outcome of the Mantel testing. The most significant correlation between genetic and geographic distances was observed among populations from the Western Mediterranean (i.e. Iberian Peninsula, South France and all the North Africans), suggesting that genetic differentiation in this region could be explained on the basis of just IBD. However, the plot of pairwise geographic vs. genetic distances from both same-coast and opposite-coast samples showed that, within the same range of geographic distances, opposite-coast genetic distances were greater than same-coast ones (Figure 4). This observation shows that the overall positive correlation in the Mantel test is actually driven by the larger genetic differences between populations on either side of the Western Mediterranean and the smaller genetic differences between populations on the same side. The clear separation of the two data collections in the plot seems to be consistent with the Mediterranean Sea acting as a genetic barrier, as proposed in previous studies[7].

Haplotype sharing

Haplotype frequencies of the two genomic regions in the 3 major geographic areas (Ivory Coast, North Africa and South Europe) are shown in [Additional file 4]. For the F7 genomic region, the samples from North Africa were found to share 24 of their 225 inferred haplotypes (10.67%) with the Ivory Coast. However, an almost identical percentage (10.41%) of haplotype sharing with the

Table 4 Mantel test for the significance of the correlation between genetic and geographic distance matrices for different sample subsets: Old World, western Mediterranean (i.e. without Crete and Turkey), North Africa, South Europe and Iberian Peninsula

	r	p	N
Old World	0.699	0.016	14
Western Mediterranean	0.478	0.005	11
North Africa	-0.083	0.531	5
South Europe	0.105	0.285	8
Western Europe	0.119	0.321	6

r: correlation coefficient, p: p-value, N: number of populations in each comparison.

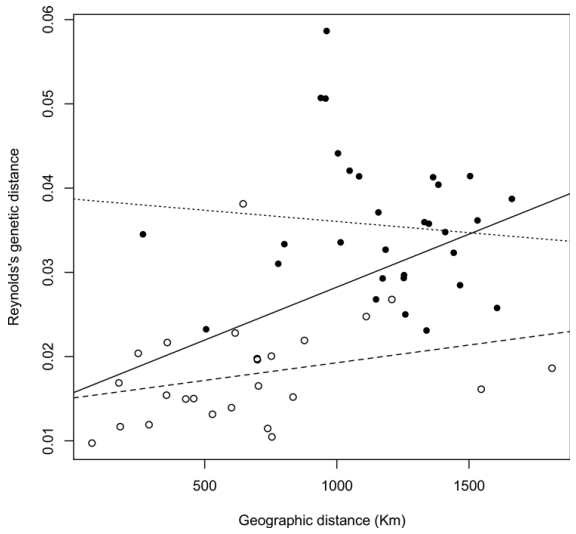


Figure 4 Scatterplot of pairwise geographic vs. genetic distances of the 11 Western Mediterranean samples. The white dots correspond to pairs of populations from the same side of the Mediterranean (i.e. NW Africa vs. NW Africa and SW Europe vs. SW Europe), while the black dots correspond to NW Africa vs. SW Europe pairs of populations. The plot reveals different distributions for the two datasets, also reflected in the respective least squares regression lines (dashed for the same-coast and dotted for the opposite-coast pairs of populations). The continuous black line represents the overall correlation.

Ivory Coast was found for South Europe (28 out of 269 haplotypes). When each sample was examined separately, the South European samples were found to share 6.35% - 16.33% of their inferred haplotypes (SNPs and microsatellites included) with the Ivory Coast, while the respective values for North Africans ranged between 8.33% and 18% [Additional file 5]. A Mann-Whitney test showed no significant differences between the two groups of values (data not shown). The same results were obtained when haplotypes based only on SNPs were used (data not shown). Since both shores of the Mediterranean share the same percentage of F7 haplotypes with the Ivory Coast, this genomic region does not provide any support to the hypothesis that the Mediterranean Sea obstructs sub-Saharan gene flow.

As for the F12 genomic region, 21.85% of the North African inferred haplotypes (33 out of 151) are shared with the Ivory Coast, while in South Europe this percentage falls to 13.30% (25 out of 188), suggesting higher gene flow from sub-Saharan Africa to North Africa. When each population was examined separately, the North African samples were found to share 20.34% - 28.07% of their inferred haplotypes (SNPs and microsatellites included) with the Ivory Coast, while the South European samples were found to share 12.20% - 22.22% of their inferred haplotypes with the Ivory Coast [Additional file 5]. In this case, the Mann-Whitney test showed that the above percentages in North African samples are significantly greater than those in South Europe (data not shown). However, no significant differences were found when haplotypes based only on SNPs were used (data not shown).

Although some of the above observations may indicate a higher gene flow from sub-Saharan Africa to North Africa than to South Europe, there is clearly no agreement between the two genomic regions or even between marker sets using both SNPs and microsatellites or just SNPs. Interestingly, a recent study based on autosomal Alu polymorphisms and compound Alu/microsatellite systems showed that gene flow through the Sahara was different between the two Mediterranean shores for the same sample set[28].

Migrant exchange estimates

Migrant exchange rates for all pairs of populations as reflected by the Nm values are shown in [Additional file 6]. As seen in the first column, the range of the Nm values for each North African sample with the Ivory Coast is 2.316 - 4.887, while the same range for the South Europeans is 1.685 - 2.627. A Mann-Whitney test showed that the Nm values of the North African samples with the Ivory Coast are significantly higher than those for South Europe with the Ivory Coast (data not shown). This finding suggests that sub-Saharan gene flow, albeit of the same order of magnitude towards both sides of the Mediterranean, is higher towards North Africa, in agreement with the data based on the F12 haplotypes presented above. Since our analyses also suggest that the Mediterranean Sea acted as a barrier to gene flow between South Europe and North Africa (see Figure 4), the differences in sub-Saharan gene flow could be interpreted as a further consequence of this barrier. However, as the significant geographic-genetic correlation for the overall Old World sample set revealed, we cannot overlook that such differences in gene flow might as well be explained on the basis of geographic distance alone.

Regarding population structure in North Africa, the Nm values tended to infinity for most pairs of North African samples, indicating the lack of any barrier to migration in the region. The only exception are the High Atlas Moroccans (Asni), which show a low rate of gene flow as compared to the remaining North African samples in agreement with the extreme position of this population in the PC plots (Figure 3).

Conclusions

With the exception of Tunisia in Figure 3a, South Europe and North Africa were always in separate clusters (Figure 3). Population structure between the two

Mediterranean shores is also supported by the results of the hierarchical AMOVA, which point towards a low but significant genetic differentiation, confirming the results of previous independent studies[7,8,27]. Moreover, the F7 and F12 variation in the Western Mediterranean presents a distribution potentially compatible with the existence a genetic barrier. However, the extent to which this is true for the whole Mediterranean could not be shown here, as for such a purpose data from other important geographic regions would be necessary such as the Italian peninsula, the Adriatic Sea (e.g. Croatia and Albania) and the Northeast Africa. Finally, the data showed no consensus regarding sub-Saharan gene flow into the two sides of the Mediterranean, thereby weighing down any evaluation of its role in the North Africa vs. South Europe differentiation. The role of the Mediterranean Sea as a barrier to gene flow is still an open case.

Additional file 1: Population allele frequencies (second row for each SNP) and heterozygosities (third row and in italics) of the 5 SNPs from the F12 genomic region. The first row for each SNP shows number of individuals typed. Polymorphisms are listed in the same order they are located on the chromosome towards the telomere. The featured frequencies correspond to the allele in bold.

Additional file 2: Population allele frequencies of the 3 microsatellite loci from the F12 genomic region. Population allele frequencies of the 3 microsatellite loci from the F12 genomic region.

Additional file 3: Variation statistics of the 3 novel microsatellite loci from the broader F12 genomic region. Total heterozygosity (H) refers to the heterozygosity of the pooled sample.

Additional file 4: Haplotype frequencies per geographic area (South Europe, North Africa, Ivory Coast) from the F7 and the F12 genomic regions. Sheet 'F7': Haplotype frequencies per geographic area from the F7 genomic region. Microsatellite loci are separated by hyphens while SNPs and the 10 bp insertion/deletion polymorphism are not separated from each other. Numbers in microsatellite loci correspond to the repeat number. SNP and INDEL alleles are annotated by numbers '1' and '2'. The 14 polymorphisms are listed in the same order they appear on the chromosome (here seen in yellow background). Sheet 'F12': Haplotype frequencies per geographic area from the F12 genomic region. Microsatellite loci are separated by hyphens while SNPs are not separated from each other. Numbers in microsatellite loci correspond to repeat number, while SNP alleles are annotated by numbers '1' and '2'. The 8 polymorphisms are listed in the same order they appear on the chromosome (here seen in yellow background).

Additional file 5: Pairwise number of shared haplotypes from the F7 and F12 genomic regions. Sheets 'F7 all markers' & 'F7 only SNPs': Pairwise number of shared haplotypes from the F7 genomic region. In sheet 'F7 all markers' estimations were based on SNPs and microsatellites, while in sheet 'F7 only SNPs' estimations were based on SNPs only. Numbers in brackets correspond to the number of distinct haplotypes found in each population. Sheets 'F12 all markers' & 'F12 only SNPs': Pairwise number of shared haplotypes from the F12 genomic region. In sheet 'F12 all markers' estimations were based on SNPs and microsatellites, while in sheet 'F12 only SNPs' estimations were based on SNPs only. Numbers in brackets correspond to the number of distinct haplotypes found in each population.

Additional file 6: Nm estimates per pair of Old World populations based on data from both the F7 and F12 genomic regions. Nm estimates per pair of Old World populations based on data from both the F7 and F12 genomic regions.

Acknowledgements
This research has been financially supported by the CGL2005-03391 and CGL2008-03955 projects of the Spanish Ministerio de Educación y Ciencia, as well as the 2005SGR00252 project of the Generalitat de Catalunya. The work of GA has been financed by an FPU grant from the Ministerio de Educación y Ciencia (grant reference: AP2005-4425). We are grateful to all of the donors for providing blood samples and to our sampling collaborators: F Luna, C Rodríguez and M De Grado (samples from Spain), N Moschonas (Crete), H Chaabani (Monastir), M Kandil and N Harich (Khenifra), M Cherkaoui (Asni), M Melhaoui (Bouhria), and N Bissar-Tadmouri (Istanbul), A Cambon-Thomsen and MS Issad (M'zab), A Chaventré (Ivory Coast) and finally M Villena (Bolivian Aymaras and Quechuas). We also want to thank the technical and material support received by the Dr. M Stoneking's laboratory of Molecular Anthropology at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, where most of the microsatellite genotyping was carried out.

Author details
[1]Unitat d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain. [2]CNRS and University Toulouse III Paul Sabatier, Toulouse, France. [3]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Authors' contributions
GA carried out the microsatellite genotyping, participated in the preparation of the DNA aliquots for SNP-typing and the statistical analyses and wrote the original manuscript. EGP carried out part of the preparation of the DNA aliquots. EE gave important advice for the improvement of the manuscript. JMD provided a substantial number of population samples to the study. MS carried out the F_{ST} comparisons and language corrections and gave substantial advice for the improvement of the manuscript. PM designed and coordinated the study and participated in the draft of the manuscript. All authors read and approved the final manuscript.

Received: 16 November 2009 Accepted: 27 March 2010
Published: 27 March 2010

References

1. Norwich JJ: *The Middle Sea: A History of the Mediterranean* London: Vintage 2007.
2. Tomas C, Sanchez JJ, Barbaro A, Brandt-Casadevall C, Hernandez A, Ben Dhiab M, Ramon M, Morling N: **X-chromosome SNP analyses in 11 human Mediterranean populations show a high overall genetic homogeneity except in North-west Africans (Moroccans).** *BMC Evol Biol* 2008, **8**:75.
3. Quintana-Murci L, Veitia R, Fellous M, Semino O, Poloni ES: **Genetic structure of Mediterranean populations revealed by Y-chromosome haplotypes analysis.** *Am J Phys Anthropol* 2003, **121**:157-171.
4. Latini V, Sole G, Doratiotto S, Poddie D, Memmi M, Varesi L, Vona G, Cao A, Ristaldi MS: **Genetic isolates in Corsica (France): linkage disequilibrium extension analysis on the Xq13 region.** *Eur J Hum Genet* 2004, **12**:613-619.
5. Angius A, Bebbere D, Petretto E, Falchi M, Forabosco P, Maestrale GB, Casu G, Persico I, Melis PM, Pirastu M: **Not all isolates are the same: linkage disequilibrium analysis on Xq13.3 reveals different patterns in Sardinian subpopulations.** *Hum Genet* 2002, **111**:9-15.
6. Picornell A, Gómez-Barbeito L, Tomàs C, Castro JA, Ramon MM: **Mitochondrial DNA HVRI variation in Balearic populations.** *Am J Phys Anthropol* 2005, **128**:119-130.
7. Comas D, Calafell F, Benchemsi N, Helal A, Lefranc G, Stoneking M, Batzer MA, Bertranpetit J, Sajantila A: **Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits.** *Hum Genet* 2000, **107**:312-319.
8. González-Pérez E, Via M, Esteban E, López-Alomar A, Mazieres S, Harich N, Kandil M, Dugoujon JM, Moral P: **Alu insertions in the Iberian Peninsula and north west Africa - genetic boundaries or melting pot?** *Coll Antropol* 2003, **27**:491-500.
9. Abdennaji Guenounou B, Loueslati BY, Buhler S, Hmida S, Ennafaa H, Khodjet-Elkhil H, Moojat N, Dridi A, Boukef K, Ben Ammar Elgaaied A, Sanchez-Mazas A: **HLA class II genetic diversity in southern Tunisia and the Mediterranean area.** *Int J Immunogenet* 2006, **33**:93-103.

10. de Maat MP, Green F, de Knijff P, Jespersen J, Kluft C: **Factor VII polymorphisms in populations with different 'risk's of cardiovascular disease.** *Arterioscler Thromb Vasc Biol* 1997, **17**:1918-1923.
11. Endler G, Exner M, Mannhalter C, Meier S, Ruzicka K, Handler S, Panzer S, Wagner O, Quehenberger P: **A common C–>T polymorphism at nt 46 in the promoter region of coagulation factor XII is associated with decreased factor XII activity.** *Thromb Res* 2001, **101**:255-260.
12. Athanasiadis G, Esteban E, Gayà-Vidal M, Dugoujon JM, Moschonas N, Chaabani H, Bissar-Tadmouri N, Harich N, Stoneking M, Moral P: **Different evolutionary histories of the Coagulation Factor VII gene in human populations?** *Ann Hum Genet* 2010, **74**:34-45.
13. Kanaji T: **Lower factor XII activity is a 'risk' marker rather than a 'risk' factor for cardiovascular disease: a rebuttal.** *J Thromb Haemost* 2008, **6**:1053-1054.
14. Belkhir K, Borsa P, Goudet J, Chikhi L, Bonhomme F: **GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations.** Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France) 1988.
15. Excoffier L, Laval G, Schneider S: **Arlequin (version 3.0): An integrated software package for population genetics data analysis.** *Evol Bioinf Online* 2005, **1**:47-50.
16. Dieringer D, Schlötterer C: **Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets.** *Mol Ecol Notes* 2003, **3**:167-169.
17. Wright S: **The genetical structure of populations.** *Ann Eugen* 1951, **15**:323-354.
18. Mann HB, Whitney DR: **On a test of whether one of two random variables is stochastically larger than the other.** *Ann Math Stat* 1947, **18**:50-60.
19. Chessel D, Dufour AB, Thioulouse J: **The ade4 package-I: one-table methods.** *R News* 2004, **4**:5-10.
20. Mantel N: **Detection of disease clustering and a generalized regression approach.** *Cancer Res* 1967, **27**:209-220.
21. Reynolds J, Weir BS, Cockerman CC: **Estimation of the coancestry coefficient: basis for a short term genetic distance.** *Genetics* 1983, **105**:767-779.
22. Felsenstein J: **PHYLIP – Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5**:164-166.
23. Stephens M, Donnelly P: **A comparison of bayesian methods for haplotype reconstruction from population genotype data.** *Am J Hum Genet* 2003, **73**:1162-1169.
24. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68**:978-989.
25. Wright S: *Evolution and the Genetics of Populations II, The Theory of Gene Frequencies* Chicago: University of Chicago Press 1969.
26. Kanaji T, Okamura T, Osaki K, Kuroiwa M, Shimoda K, Hamasaki N, Niho Y: **A common genetic polymorphism (46 C to T substitution) in the 5'-untranslated region of the coagulation factor XII gene is associated with low translation efficiency and decrease in plasma factor XII level.** *Blood* 1998, **91**:2010-2014.
27. Athanasiadis G, Esteban E, Via M, Dugoujon JM, Moschonas N, Chaabani H, Moral P: **The X chromosome Alu insertions as a tool for human population genetics: data from European and African human groups.** *Eur J Hum Genet* 2007, **15**:578-583.
28. González-Pérez E, Esteban E, Via M, Gaya-Vidal M, Athanasiadis G, Dugoujon JM, Luna F, Mesa MS, Fuster V, Kandil M, Harich N, Bissar-Tadmouri N, Saetta A, Moral P: **Population relationships in the Mediterranean revealed by autosomal genetic data (Alu and Alu/STR compound systems).** *Am J Phys Anthropol* 2009, **141**:430-439.

doi:10.1186/1471-2148-10-84
Cite this article as: Athanasiadis *et al.*: The Mediterranean Sea as a barrier to gene flow: evidence from variation in and around the F7 and F12 genomic regions. *BMC Evolutionary Biology* 2010 **10**:84.

Discussion

OVERALL DISCUSSION OF THE RESULTS

In this work, a variety of genetic markers were used in order to address different issues regarding the genetic profile of the Mediterranean, as well as that of other regions (i.e. sub-Saharan Africa and Native American Bolivia). To do so, 13 Alu polymorphisms from the X chromosome were analysed for the first time in six populations (525 individuals) from the broader Mediterranean region and sub-Saharan Africa (Athanasiadis et al., 2007). In addition, variation from two autosomal genomic regions, related to the risk of cardiovascular disease, was analysed in 687 individuals from 16 worldwide populations – 13 of them located in the broader Mediterranean region. The following paragraphs expose the main findings of this work.

On the differentiation between North Africa and South Europe

Undoubtedly, one of the main topics of this work is the genetic differentiation between North African and South European populations. As most of our data suggested, this differentiation is low but significant.

The North Africa – South Europe differentiation is most noticeable in the three PC plots from the survey of population structure based on the variation from the F7 and F12 genes (Athanasiadis et al., 2010b): in all three PC plots, the 13 samples from the broader Mediterranean region were always significantly clustered into two groups – North Africa and South Europe. Notwithstanding, this differentiation was not as obvious for the X chromosome Alu insertions (Athanasiadis et al., 2007) most probably due to the low number of samples considered and some particularities that, as we shall see further down, the

Tunisian sample presents. In any case, the principal component analysis indicated the clear separation of the North African Berbers (Morocco and Egypt) from the European samples (see also Appendix 1).

The North Africa – South Europe differentiation is also evident in the results from the analysis of variance of the allelic frequencies, where the autosomal markers from the F7 and F12 genomic regions (Athanasiadis et al., 2010b) indicated that 1.34% of the genetic variance found in the Mediterranean samples was attributable to the South Europe vs. North Africa clustering (F_{CT} = 0.013, $p<0.05$). Our results are consistent with previous studies of population relationships in the Western Mediterranean based on autosomal Alu polymorphisms (Comas et al., 2000; González-Pérez et al., 2003) and Alu/microsatellite compound systems (González-Pérez et al., 2009), which reported a significant – although somewhat higher – differentiation between the two shores (F_{CT} values: 0.020, 0.018 and 0.022 respectively).

Contrary to the significant differentiation found when autosomal markers were examined, the hierarchical AMOVA based on the X chromosome Alu polymorphisms (Athanasiadis et al., 2007) showed that the low differentiation between the two sides of the Mediterranean Sea was not significant (F_{CT} = 0.012, $p>0.05$). This lack of significance could be explained by the lower number of samples considered in that study, although the particularity of the X chromosome could also be relevant to our observations; since males have only one copy of the X chromosome, every existing X chromosome has spent two-thirds of its history in females. This fact has two interesting consequences. First, genetic diversity is lower in the X chromosome than in the autosomes, because

the nucleotide mutation rate in females is several times lower than in males (Schaffner, 2004). This low diversity is exactly what we observed in our data. Second, the X chromosome polymorphisms mainly reflect female history. It has been proposed that the Mediterranean Sea presents a higher permeability to female than to male migration (Plaza et al., 2003; Tomas et al., 2008). As a consequence, the X chromosome, being more representative of the female history, is expected to present a looser population structure as compared with the autosomes, which once again is exactly what we observed in our data.

The low and nonsignificant differentiation between the two Mediterranean shores based on the X chromosome Alu polymorphisms is also confirmed by the STRUCTURE analysis of the same data (Athanasiadis et al., 2007). When the five Mediterranean groups were analysed, STRUCTURE showed that the most plausible model to fit the data was that of K=1 ancestral population, suggesting the lack of a differentiation important enough to classify the samples into more than one cluster.

In general, most of the markers used in this study, with the exception of the X chromosome, are consistent with a low but significant genetic differentiation between populations from the southern and northern Mediterranean coasts.

On the Mediterranean Sea as a genetic barrier

The low but significant differentiation between North Africa and South Europe suggests that the Mediterranean Sea has been (at least partially) a barrier to gene flow. A genetic barrier imposed by the Mediterranean Sea is expected to obstruct gene flow not only between North Africa and South Europe, but also

between sub-Saharan Africa and South Europe. As a consequence, sub-Saharan gene flow would be less intense towards South Europe than towards North Africa.

Some support to this hypothetical situation can be found in the haplotype analysis of the F12 genomic region, as well as in the migration patterns represented by the overall Nm values (Athanasiadis et al., 2010b). Both on a broader geographical and on a single population level, the North Africans were found to share significantly more F12 haplotypes with the Ivory Coast than the South Europeans did. Moreover, the analysis of migration patterns based on both the F7 and the F12 genomic regions showed that sub-Saharan gene flow – as reflected by the Nm values – was significantly higher towards North Africa than it was towards South Europe.

However, different results were obtained from the haplotype analysis of the F7 genomic region: neither on a broader geographic nor on a single population level were the North Africans found to share a significantly higher percentage of F7 haplotypes with the Ivory Coast as compared with the South Europeans. Besides, the significant matrix correlation between genetic and geographic distances of the Mediterranean and sub-Saharan samples suggests that the haplotype patterns found could well be attributed to isolation by distance alone.

In view of the controversial results from the F7 and F12 genomic regions, we found in the literature a recent study which showed that sub-Saharan gene flow was different between the two Mediterranean shores (González-Pérez et al., 2009), providing further support to the idea of the Mediterranean as a barrier to gene flow. While this issue is still open for debate, it is important to remember

that, since we are dealing with autosomal markers, all the above affirmations are based on inferred – not unequivocal – haplotypes.

Regarding the extent of the genetic barrier, previous studies locate it in the Western Mediterranean and more specifically in the Gibraltar Strait (Comas et al., 2000; Bosch et al., 2001). Interestingly, the genetic and geographic distances among the Western Mediterranean populations of our study were highly correlated (Athanasiadis et al., 2010b). This observation suggests that isolation by distance might have actually operated as a mechanism of population differentiation in the area, a scenario that is not necessarily compatible with the existence of a genetic barrier. However, the plot of geographic vs. genetic distances was consistent with the existence of a genetic barrier in the Western Mediterranean (Athanasiadis et al., 2010b), also supported by other studies in which abrupt frequency changes between Northwest Africa and the south of the Iberian Peninsula were reported (Bosch et al., 2001, Malaspina et al., 2000).

A further investigation of the issue based on a matrix correlation analysis in the Eastern Mediterranean was not possible, as this would not be reliable due to the poor sampling of this region. Notwithstanding, evidence for the lack of a genetic barrier in the Eastern Mediterranean can be found in a recent study, which reported direct trans-Mediterranean migrations from North Africa to Europe based on the analyses of a Y chromosome haplogroup (Cruciani et al., 2007).

In conclusion, it seems plausible that gene flow through the Mediterranean Sea has been, at least partially, impaired. However, the genomic variation analysed in this study does not provide enough information for a more detailed inquiry into

this controversial issue; more data are necessary to explore the role of sub-Saharan gene flow in the differentiation between the two Mediterranean Coasts.

On the genetic structure of the South Europeans

In general terms, our data suggest that South Europe does not fit a pattern of strong genetic structure. This is most obvious in the three PC plots from the study of the F7 and F12 genomic regions (Athanasiadis et al., 2010b), where no significant clustering in geographical terms seems to arise for the South European samples. Moreover, the hierarchical AMOVA based on this autosomal marker set indicated a low and nonsignificant differentiation among the South European populations, reflected in the F_{ST} values (South Europe: F_{ST}=0.0004, $p>0.05$). Conversely, for the X chromosome Alu polymorphisms, the hierarchical AMOVA revealed higher and significant F_{ST} values in South Europe (F_{ST} = 0.004, $p<0.001$) that may be related to the lack of a systematic east-west female migration and/or faster genetic drift of the X chromosome (Schaffner, 2004).

Other recent studies of population structure based on gradually growing SNP datasets from the whole genome showed a rough north – south/southeast differentiation within the European continent (Seldin et al., 2006; Bauchet et al., 2007; Tian et al., 2008). In line with our results, these studies fail to capture a strong genetic structure in their South European samples, with the notable exception of the Spanish samples in Bauchet et al., 2007. However, more recent studies based on several hundred thousand SNPs found a strong correlation between genes and geographic origin within the entire European terrain, although low average levels of genetic differentiation among Europeans were generally admitted (Novembre et al., 2008; Nelis et al., 2009). It therefore

becomes clear that subtle population structure in South Europe is expected to arise more easily when high-volume genetic data are available than when a number of markers as low as that of our studies is considered.

Additionally to these general considerations about South European population structure, the genetic position of the Basques in the South European genetic landscape deserves a mention. In general, our markers are in agreement with other studies, which found no special genetic features for the Basques.

The X chromosome Alu insertions placed the Basques close to the other European sample, the Cretans (Athanasiadis et al., 2007). It is also worth noting that, in the same study, the only genetic distance that was not statistically different from zero was that of the Basque Country with the island of Crete. The close distance between these two groups can also be seen in the lack of any significant differences between them when a locus-by-locus χ^2 comparison was performed. Similarly, in all the PC plots that appear in the study of the F7 and F12 genomic regions (Athanasiadis et al., 2010b), the Basques were always plotted close to Pas Valley, Catalonia or South France. Even more interestingly, in the same study, most of the Nm values between the Basques and the other samples from the Iberian Peninsula and South France tended to infinity, reflecting high migration activity – i.e. the lack of any genetic barriers.

From all these observations, it seems quite reasonable to argue that the Basques – firmly located within continental Europe and with no impenetrable geographical barriers surrounding them – are in general terms no more genetically special than most of the European populations. This conclusion is also supported by data from other studies – mostly based on DNA markers –

carried out throughout recent years (e.g. Comas et al., 2000; Harich et al., 2002; Alonso et al., 2005; García-Obregón et al., 2007; Garagnani et al., 2009). Nevertheless, the fact that the Basques do present unusual allele frequencies for several marker systems (Bauduer et al., 2005) reminds us that the debate is still open.

On the genetic structure of the North Africans

Unlike South Europe, our data revealed a more accentuated population structure in North Africa.

Starting with the Berbers from Egypt (Siwa Oasis), this was the only North African sample in our studies that did not come from the Maghreb (Northwest Africa). It is therefore important to pinpoint their genetic position in the North African landscape. The Siwa Oasis group has been previously described as a genetic outlier (Esteban et al., 2006; Moral et al., 2006; Coudray et al., 2009) due to the desert surrounding it. Some evidence of this observation can be found in the PC plot based on the X chromosome Alu insertions (Athanasiadis et al., 2007), where Siwa Oasis is in fact separated from both the Ivory Coast and the rest of the Mediterranean samples. Moreover, Siwa Oasis is the North African group with the closest genetic affinity with the Ivory Coast, as the Reynolds' distances showed. This relationship between the Egyptian Berbers and the Ivory Coast could be attributed to a greater sub-Saharan gene flow into Siwa Oasis through the Nile River, as data from Alu/STR systems (González-Pérez et al., 2009) and the variation of a trinucleotide in the gene of the androgen receptor (Esteban et al., 2006) revealed recently.

Moving from Egypt to the Maghreb, our attention is first drawn to the Arab-speaking sample from Tunisia, which seems to occupy a particular genetic position with respect to the other North African groups. The genetic distances between Tunisia and the two European samples in the study of the X chromosome Alu insertions (Athanasiadis et al., 2007) are shorter than those between Tunisia and the other North African groups, also reflected in the PC plot from the same study. Tunisia is also plotted closer to the South Europeans in the PC plot of the overall sample set from the study of the F7 and F12 genomic regions (Athanasiadis et al., 2010b). This observation is in agreement with a recent study, in which Tunisians did not show a significant level of differentiation from South European populations (Tomas et al., 2008).

As the mutation rate of the Alu elements is generally low, the affinity between Tunisia and South Europe could be reflecting older aspects of the Mediterranean genetic landscape, such as a common pre-Neolithic population stock on both North and South Mediterranean shores.

However, another possible explanation for the special genetic position of this Arab-speaking Tunisian sample can be found in the genetic composition of the region: Tunisia presents a considerable genetic heterogeneity (Giraldo et al., 2001; Fadhlaoui-Zid et al., 2004), whereby the country is inhabited by Arabs and Berbers, but also by people of other ethnic/cultural backgrounds, like Jews and dark-skinned people of sub-Saharan origin (Ennafaa et al., 2006). As a consequence, different Tunisian samples are expected to present different degrees of genetic affinity with South European and other North African groups.

Unlike the X chromosome Alu polymorphisms, the study of the F7 and F12

genomic regions (Athanasiadis et al., 2010b) indicated that the migration rate (reflected by the Nm values) was generally low between Tunisians and South Europeans and high between Tunisians and the rest of the North Africans. According to these markers, the Tunisian sample has greater genetic affinity with the North African samples than with the European ones. This observation is consistent with previous studies, in which a Tunisian sample representing the entire country did not exhibit any significant differences from other North African samples (Bahri et al., 2008). As a result of the contradiction between the results from the X-chromosome and autosomal markers, no definite conclusion about the genetic relationships of Tunisia with South Europe and other North African regions can be reached.

Moving further to the west, the Moroccan samples from the High Atlas seem to present some noteworthy features; in the study of the F7 and F12 genomic regions (Athanasiadis et al., 2010b), the High Atlas Moroccans (Asni and Khenifra) were clustered far from the other North Africans. Moreover, in the X chromosome study (Athanasiadis et al., 2007), where only the Asni Moroccans were examined, they presented an outlying position in relation with other North African groups. These observations could be correlated with a previous study, which reported the differentiation of a Moroccan sample from other Mediterranean groups (Tomas et al., 2008).

Various hypotheses have been put forward in order to explain the differentiation in Northwest Africa. Several studies speculate an early settlement of Northwest Africa or an early genetic drift – that could have driven allele frequencies to highly differentiated values – combined with a low gene flow into the region

from the surrounding areas (Tomas et al., 2008; Arredi et al., 2004; Bosch et al., 2000). According to this hypothesis, the isolation of Northwest Africa was mainly geographical, with the Strait of Gibraltar in the north (Comas et al., 2000; Bosch et al., 2001) and the Sahara desert in the south (Arredi et al., 2004) playing the role of genetic barriers.

All the same, the differentiation of the High Atlas could be related to a stronger genetic influence from sub-Saharan Africa. Indeed, gene flow – as reflected by the Nm values – between the Ivory Coast and Northwest Moroccans was found to be higher than it was between the Ivory Coast and the other North Africans in the pertinent study (Athanasiadis et al., 2010b). This hypothesis opposes the idea of an isolated Northwest Africa mentioned above. Besides, a certain level of genetic exchange is thought to have occurred between NW Africa and the south of the Iberian Peninsula (Bosch et al., 2001; Kandil et al., 1999; Scozzari et al., 2001; Malaspina et al., 2000).

In conclusion, our observations regarding Siwa Oasis, Tunisia and the Northwest Moroccans indicate that North Africa presents a considerable population structure as compared with South Europe. This fact is likely to be due to the role of the Sahara in the isolation of certain populations and the different effect of sub-Saharan gene flow on North African groups.

On the evolutionary history of the F7 genomic region

The diverse results considering positive selection in our samples suggest that the F7 promoter may have undergone different evolutionary paths in different parts of the world. More specifically, the 'risk' vs. neutral F_{ST} comparisons and the long-

range haplotype test produced similar results for the Mediterranean region and the Ivory Coast and a different result for the Native Americans.

The F_{ST} comparisons in the three sample sets (i.e. All Populations, Old World, Mediterranean) did not yield any 'risk' F_{ST} values significantly greater than the neutral F_{ST} values, as one would expect under strong positive selection (Athanasiadis et al., 2010a). This observation is in agreement with a previous study where a similar F_{ST} method was used to reveal no evidence for positive selection for polymorphisms −324, −401 and −402 in an Italian sample (Hahn et al., 2004).

Similarly, the long-range haplotype analysis in the same study (Athanasiadis et al., 2010a) showed that no signal of positive selection for the F7 promoter could be found in South Europe, North Africa or the Ivory Coast. On the other hand, the Native Americans presented quite a different picture with the higher 'risk' core haplotype presenting both the highest frequency as well as the highest extended haplotype homozygosity – a typical sign of positive selection (Sabeti et al., 2002).

In the absence of data on functional and phenotypic differences among different F7 variants (which could provide a solid proof of positive selection in some of our populations), no other choice is left but to rely on theoretical expectations; the lack of any signal of positive selection around the Mediterranean suggests that in this geographical region the variation in the F7 promoter did not bring any selective advantages to the potential carriers and that this variation is most probably the result of random genetic drift. In this sense, the F7 'risk' markers in the Mediterranean samples are essentially no different from the neutral markers,

a fact that was exploited in most of the population analyses (Athanasiadis et al., 2010a; 2010b).

As for the Native Americans, the F_{ST} comparisons showed that the remarkably high F_{ST} values of the −670C, −630G and −402A variants were also better explained by genetic drift (i.e. no significant differences were present between neutral and 'risk' markers). However, the long-range haplotype test – which is more sensitive to recent selective events (Sabeti et al., 2002) – revealed an unusual pattern for the 'risk' core haplotype in the Native Americans. As already mentioned, this observation alone cannot be interpreted as a solid proof of selection; besides, there is a reasonable possibility that EHH/frequency differences in Native Americans can be attributed to drift.

In any case, the possibility of positive selection in the F7 promoter as a response to a more efficient recovery after injury (Lindqvist and Dahlbäk, 2008) is still open with regard to the Native American populations. In fact, signals of positive selection were found in other populations of Asian origin, such as a Chinese sample from Singapore (Hahn et al., 2004) and a Yakut sample from the Sakha Republic in Siberia (see Appendix 3).

These observations lead to the conclusion that the F7 gene may actually be reflecting different evolutionary histories in different parts of the world, according to the following scheme: transition G>A at locus −402 leads to higher FVII plasma levels and to more efficient blood clotting after a haemorrhage. However, in the Mediterranean region, this increased efficiency was either indifferent or cancelled out by other unknown factors, as the lack of positive selection indicates. On the other hand, evidence of a signal of positive selection in

some populations of Asian descent indicates that the same polymorphism under different circumstances might have had a considerable effect on fitness. To further investigate this hypothesis, more genetic data from other loci involved in blood coagulation are necessary, as well as more epidemiological data on cardiovascular incidence and historical data on the living conditions in different parts of the world.

On the role of FXII 46C>T in the risk of ischemic heart disease in the Western Mediterranean

The role of the 46C>T polymorphism in the risk of ischemic heart disease was investigated in the context of a more general screening of several candidate regions for cardiovascular disease, with the intention of 'confirmation by replication'. Our data did not reveal any role of this polymorphism as a risk factor of ischemic heart disease in patients from either Barcelona or Tunisia.

Considering the family study from Barcelona (Athanasiadis et al., 2009), the TDT revealed no significant transmission of the 'risk' allele T from the parents to the patients ($p>0.05$). However, our sample lacked enough statistical power to detect a relative risk as low as the one found in this study (1.17); the minimum relative risk our 101 family trios could possibly detect, assuming a statistical power of >80%, was 1.90. This means that more family trios would be necessary in order to detect a relative risk as low as 1.17. As this is the first time – to our knowledge – that polymorphism 46C>T has been tested through a family-based TDT, more family trios should be tested in the future in order to reach a solid conclusion. In this context, we believe that our data could prove quite useful in future meta-analyses.

As for the Tunisian case-control study, this returned a low and nonsignificant odds ratio for the T/T genotype, supposedly associated with ischemic heart disease (OR=1.36, p>0.05). Again, the minimum odds ratio our sample of 76 patients and 118 controls could possibly detect was approximately 1.90 (assuming a statistical power of >80%), so our sample was not statistically powerful enough to detect an odds ratio as low as 1.36. However, the odds ratio found in our study is much lower compared with a reported odds ratio of 4.8 (Santamaría et al., 2004). Since our Tunisian sample would have enough power (>99%) to detect an odds ratio as high as 4.8, we believe that the detected lack of association in our study points in the right direction.

The contradictory results obtained in the studies of association between polymorphism 46C>T and cardiovascular disease (Kanaji et al., 2008; Bach et al., 2008) could generally be reflecting the lack of a causal relationship between these two features. While it is well established that the T allele decreases the FXII plasma levels in a predictable way (Kanaji et al., 1998), patients of all three phenotypes (C/C, C/T and T/T) were found to have significantly lower plasma levels than the controls with the same genotypes (Kanaji et al., 2006). The above observations together support the hypothesis that FXII itself is neither a cause nor a risk factor for ischemic heart disease, and that lower FXII plasma levels are most probably a consequence of thrombosis. Unfortunately, the lack of FXII plasma level measurements from our study makes any similar assessment impossible.

To sum up, our study – far from proposing causal mechanisms – revealed a weak lack of association between polymorphism 46C>T and ischemic heart disease,

thereby supporting the idea that this polymorphism is not a universal independent risk factor for cardiovascular diseases. However, the low sample size of the family study, the unmatched structure between cases and controls in the Tunisian sample and the lack of plasma measurements demand some caution in the interpretation of our results.

Conclusions

CONCLUSIONS

1. The variation of the genomic regions analysed indicates that the genetic differentiation between North African and South European populations is low but significant.
2. The Mediterranean Sea may have at least partially acted as a barrier to gene flow between North Africa and South Europe, most likely to be located in Western Mediterranean.
3. Sub-Saharan gene flow towards the two Mediterranean Coasts might have been affected by the existence of a genetic barrier in the Mediterranean, but the possibility that the observed differentiation patterns are due to isolation by distance alone cannot be discounted, as the significant correlation between matrices of genetic and geographic distances showed.
4. The genetic differentiation in the European side of the Mediterranean region is less pronounced than it is in the African side. North Africa presents a certain degree of population structure that is evident even when low numbers of markers are studied.
5. In contrast with previous studies, whereby some populations were traditionally considered to be genetic isolates, the genetic variation studied in this work showed that the Basques do not present any special genetic position with respect to other South European populations. This finding is in agreement with other recent studies.
6. The Egyptian Berbers from Siwa Oasis appear as the most differentiated population within North Africa presenting considerable sub-Saharan

influences, likely to be reflecting an important geographic isolation imposed by the Sahara desert.

7. Within North Africa, the Tunisian population appears as the genetically closest to Europe when the X chromosome variation was considered, possibly due to the historical background of the region.
8. The genetic variation from both the X chromosome and the autosomal markers showed that the Berbers from Northwest Africa (Asni and Khenifra) were highly differentiated from the rest of the North Africans, possibly owing to a higher sub-Saharan genetic influence.
9. The population distribution of the variation in the F7 promoter region does not present any signs of positive selection in the Mediterranean region. Conversely, there is some solid evidence of positive selection in the Native American populations from Bolivia.
10. In contrast with several previous studies, the lack of association between polymorphism FXII 46C>T and ischemic heart disease in the statistically robust Tunisian case-control design suggests that this polymorphism is not a universal risk factor of ischemic heart disease.

Resumen en castellano

INTRODUCCION

Variación genética y genética de poblaciones humanas

Los seres humanos varían en casi todas sus características. Una gran parte de esta variación es genética y bien puede tener una manifestación observable o no. La variación genética es la materia prima de la genética de poblaciones humanas.

Cada fracción de la variación genética refleja distintos aspectos de la evolución humana: por un lado, la variación neutra proporciona un registro pasivo de la historia demográfica humana. Por otro lado, la variación funcional, con sus consecuencias biológicas y evolutivas, desempeña un papel fundamental en la comprensión de otro aspecto de la historia de la humanidad, el biológico, descubriendo importantes aspectos de los procesos genéticos que moldearon nuestra especie.

La variación genética observada en las poblaciones humanas actuales es el resultado de cinco procesos evolutivos: mutación, recombinación, deriva genética, selección y migración. La mutación genera nuevos alelos, mientras que la recombinación produce nuevas combinaciones de la variación genética existente. Por otro lado, los otros tres procesos actúan como modificadores de frecuencias alélicas a nivel poblacional: un alelo nuevo puede ser eliminado o aumentar en frecuencia a través de procesos estocásticos (deriva genética). Asimismo, los loci funcionales pueden ser sometidos a un proceso selectivo en función de los efectos que ellos tienen sobre la supervivencia y la capacidad reproductiva de sus portadores. Por último, la migración afecta la distribución geográfica de la diversidad genética humana.

Categorías de marcadores moleculares que se han utilizado en este trabajo

Inserciones Alu

Las secuencias Alu son secuencias cortas repetidas dispersas por el genoma de una longitud alrededor de unos 300 pares de bases. Con aproximadamente más de un millón de copias, las secuencias Alu son los elementos móviles más abundantes en el genoma humano. Algunas de las inserciones Alu son polimórficas en el sentido de que la inserción de la secuencia no está fijada; un cromosoma puede bien portar la inserción, bien no.

Dichos polimorfismos Alu resultan de gran utilidad para los estudios antropológicos que investigan el origen y las relaciones genéticas entre diversas poblaciones humanas, debido a dos importantes ventajas sobre otros marcadores genéticos: por un lado, los portadores de una inserción en un locus determinado son idénticos por descendencia y, por otro, el estado ancestral de un polimorfismo Alu puede ser determinado y casi siempre corresponde a la ausencia de inserción.

La variación de las secuencias Alu del cromosoma X en poblaciones humanas se ha descrito anteriormente pero dichos polimorfismos no se habían usado en estudios poblacionales, representando este trabajo uno de los primeros estudios sistemáticos sobre la variación de las inserciones Alu polimórficas del cromosoma X.

Microsatélites

Los microsatélites son perfectas o casi perfectas repeticiones en tándem de secuencias cortas, generalmente mono-, di-, tri- y tetranucleótidos, aunque muy a menudo los penta- y hexanucleótidos también se definen como microsatélites. Los microsatélites se caracterizan por una alta variabilidad y tasa de mutación reflejadas en sus múltiples alelos y alta heterocigosidad. La variación de los microsatélites consiste en el número de repeticiones en cada alelo.

Polimorfismos de un solo nucleótido (SNPs)

Los polimorfismos de un solo nucleótido (SNPs) son los más sencillos de los marcadores moleculares y consisten en la sustitución de una base nucleotídica por otra. Generalmente, los SNPs se generan tras la incorporación errónea de nucleótidos durante la replicación por puro azar o como consecuencia de una mutagénesis química o física. Sus tasas de mutación son tan bajas que los SNPs son prácticamente idénticos por descendencia: dos individuos que portan el mismo alelo en un locus lo han heredado de un antepasado común. La baja tasa de mutación también significa que los SNPs son marcadores principalmente bialélicos.

Movimientos poblacionales en el mediterráneo

El mediterráneo presenta una serie de características que lo convirtieron en uno de los escenarios más importantes de la historia de la humanidad. A continuación se presentan brevemente los eventos migratorios que más han afectado la composición genética de las poblaciones humanas del mediterráneo.

Expansiones prehistóricas

Según el registro fósil y arqueológico, la colonización de Europa por los humanos anatómicamente modernos se produjo hace unos 40.000 años. Estos primeros pobladores trajeron sus tecnologías a Europa desde el Oriente Próximo a través de los Balcanes y de allí al oeste. Semejantes dispersiones desde el Oriente Próximo se produjeron en el norte de África.

Durante el último máximo glacial (unos 19.500 – 25.000 años antes del presente), las poblaciones humanas buscaron refugio en el suroeste de Europa, a lo largo del Mediterráneo, los Balcanes y el Levante, y en las llanuras de Europa oriental. Desde estos refugios las poblaciones humanas volvieron a expandir hacia el centro y norte de Europa una vez que cesaron las circunstancias climáticas extremas.

El comienzo del Neolítico fue marcado por la llegada de la agricultura (unos 10.000 años antes de ahora) y constituye, sin duda, uno de los eventos más importantes en la historia de la humanidad. La difusión de la tecnología agrícola es uno de los debates más conocidos en el ámbito de la evolución humana. En general, hay dos modelos rivales: difusión démica vs. difusión cultural. Según el modelo de difusión démica, la expansión agrícola fue producida por la migración progresiva de los grupos agrícolas del Oriente Próximo, sustituyendo en su camino a las poblaciones nativas. En cambio, el modelo de difusión cultural sugiere que la nueva tecnología se difundió a través del intercambio y la asimilación de las nuevas tecnologías entre grupos humanos vecinos.

Según el primer modelo, a medida que los portadores de la nueva tecnología avanzaban, la diversidad genética de la mayoría de las poblaciones autóctonas debieron de ser dramáticamente afectadas. Sin embargo, se cree que algunas de estas poblaciones indígenas conservaron parte de su diversidad genética original, debido a una serie de mecanismos aisladores. Ejemplos de dichas poblaciones son los sardos, los corsos y los baleares, aunque quizás los mejor estudiados sean los vascos.

Movimientos poblacionales históricos

Con la llegada de la Edad de Bronce en el Oriente Próximo (unos 3.300 años aC), aparecen las primeras rutas comerciales en el Mediterráneo. Primero los fenicios y luego los griegos, seguidos por los romanos, establecieron colonias comerciales a lo largo de ambas costas mediterráneas. Mientras tanto, la desertificación gradual del ecosistema terrestre en el norte de África, que comenzó hace unos 6.000 años, marcó la transición de un Sahara verde al desierto hiperárido de ahora, de manera que el Sáhara puede haber operado como una barrera geográfica importante en el sur del Mediterráneo, favoreciendo el aislamiento de las tribus nativas norteafricanas (los bereberes). Sin embargo, este aislamiento nunca llegó a ser completo.

A lo largo de la historia romana, el mediterráneo fue invadido por varios pueblos procedentes del centro y norte de Europa. Un ejemplo de ello son los celtas, cuya expansión comenzó alrededor del 450 aC, llegando hasta la Península Ibérica, donde compartieron territorio con los íberos autóctonos. Asimismo, el período de la historia europea entre el 300 y 700 dC se conoce como “Las invasiones bárbaras”. Durante este período hubo una serie de consecutivas oleadas

migratorias hacia el interior del imperio romano por parte de varias tribus germánicas. Las más importantes fueron los vándalos, los godos y los lombardos.

Por último, la invasión de los árabes en el norte de África y la Península Ibérica fue uno de los últimos grandes acontecimientos migratorios que han afectado la composición genética actual de las poblaciones del mediterráneo occidental. Hacia finales del siglo VII, los árabes ya habían conquistado el Magreb bereber y estaban a punto de a cruzar el Estrecho de Gibraltar para invadir la Península Ibérica. Durante los casi 750 años de ocupación musulmana, el territorio español fue compartido entre sus habitantes antiguos (romanos y visigodos) y los recién llegados (árabes y bereberes).

La genética de poblaciones humanas en el mediterráneo

Los primeros árboles filogenéticos detallados se publicaron a finales de los años 80 y utilizaban la información genética del ADN mitocondrial y el cromosoma Y. Estas dos fracciones del genoma humano siguen utilizándose en los estudios antropológicos, incluidos los estudios mediterráneos. Además de los marcadores uniparentales, en las últimas décadas también ha crecido la investigación basada en marcadores autosómicos, un hecho que comenzó en las décadas de los 70 y 80 con el estudio de la distribución global de distintos grupos sanguíneos humanos (incluyendo el Mediterráneo) pero que, con el advenimiento de las nuevas tecnologías moleculares, ha alcanzado unas dimensiones espectaculares.

Todos estos marcadores se han utilizado para contestar a preguntas recurrentes en el campo de la antropología evolutiva. En el caso concreto del mediterráneo occidental, una pregunta de este tipo es el papel del Estrecho de Gibraltar en la

diferenciación de las poblaciones ibéricas a las poblaciones norteafricanas. Algunos especialistas consideran el Estrecho de Gibraltar como una barrera genética entre dicha poblaciones, mientras que otros lo tratan como un puente de difusión cultural y genética. Por ejemplo, el estudio de González-Pérez y colaboradores en 2003 basado en Alu autosómicas mostró que la diferenciación genética encontrada entre los dos lados del mediterráneo occidental se podría explicar con un modelo de aislamiento por distancia, un hallazgo que además encuentra apoyo en otros estudios. Sin embargo, al mismo tiempo la diferenciación genética encontrada en otros estudios sostiene que el Estrecho de Gibraltar ha operado como una barrera al flujo génico. El debate todavía está abierto.

En cuanto al campo de la epidemiología genética, existe un elevado número de estudios de la base genética de enfermedades complejas. La investigación realizada por el grupo investigador en el seno del cual se ha realizado en presente trabajo, está desarrollada principalmente alrededor de las enfermedades cardiovasculares y más concretamente en el papel de una serie de polimorfismos funcionales (procedentes de los genes de las apolipoproteínas y las sintasas de óxido nítrico) en el desarrollo de la cardiopatía isquémica en el mediterráneo.

Las enfermedades cardiovasculares

Las enfermedades cardiovasculares son generalmente las enfermedades del corazón y de los vasos sanguíneos, pero el término se usa principalmente para referirse a las que tienen un origen aterosclerótico. Por causa de la aterosclerosis, los vasos sanguíneos se constriñen y pierden su elasticidad

resultando a la obstrucción del flujo sanguíneo. El evento culminado de la aterosclerosis es la trombosis (la formación de coágulos), durante la cual el flujo sanguíneo se interrumpe con graves consecuencias.

Una de las enfermedades cardiovasculares más comunes es la cardiopatía isquémica. Dicha alteración es producida por una disminución en la proporción de sangre al miocardio, normalmente como consecuencia de aterosclerosis en las arterias coronarias. Las dos manifestaciones más habituales de la cardiopatía isquémica es la angina de pecho y el mucho más deletéreo infarto de miocardio. Sólo en 2005 las enfermedades cardiovasculares mataron alrededor de 17.5 millones de personas en todo el mundo, representando el 30% del total de fallecimientos. De estas muertes, 7.6 millones se atribuyen a la cardiopatía isquémica.

La trombosis está relacionada con la hemostasis, un complejo proceso que para la hemorragia mediante la acción sincronizada de espasmos vasculares, la formación del tapón plaquetario (hemostasis primaria) y la coagulación de la sangre (hemostasis secundaria). Durante la coagulación, los coágulos se forman mediante la conversión del fibrinógeno en fibrina como respuesta a una secuencia de activaciones enzimáticas conocida como "la cascada de coagulación". Dichas enzimas se llaman factores de coagulación, siendo la mayoría de ellos proteasas de serina que actúan mediante la rotura de otras proteínas en sitios específicos.

En este trabajo nos hemos centrado en dos factores de coagulación, el factor VII y el factor XII:

Factor de Coagulación VII

El Factor de Coagulación VII es una glucoproteína dependiente de la vitamina K que se sintetiza en el hígado y se secreta en el torrente sanguíneo como un zimógeno inactivo. Tras una herida endotelíaca, el Factor Tisular activa el Factor VII. El complejo del Factor VII activado con el Factor Tisular es el que inicia la vía intrínseca de la cascada de la coagulación. Muchos estudios epidemiológicos afirman que unos niveles plasmáticos de Factor VII altos están relacionados con un alto riesgo de enfermedades cardiovasculares.

Según un estudio, los niveles plasmáticos del Factor VII presentan un 53% de heredabilidad y se determinan por un solo locus, el del gen F7. El gen F7 ocupa unos 12.8 kb en la región cromosómica 13q34 y comprende en 9 exones y 8 intrones. La proteína madura codificada por dicho gen consiste en 406 aminoácidos.

Se ha descubierto que la variabilidad en los niveles del Factor VII depende esencialmente de variantes reguladoras no codificantes y variantes intrónicas, más que de variantes que resultan en cambios aminoacídicos.

Factor de Coagulación XII

El Factor de Coagulación XII es un precursor de una proteasa de serina producido y secretado por hepatocitos. El Factor XII está principalmente involucrado en dos caminos bioquímicos opuestos: por un lado la iniciación de la vía extrínseca de la cascada de coagulación y por otro la fibrinólisis.

El gen que codifica para el Factor de Coagulación XII (F12) ha sido objeto de muchos estudios epidemiológicos de riesgo cardiovascular, con resultados controvertidos.

Estudios previos han mostrado que los niveles del FXII en el plasma de la sangre exhiben un 67% de heredabilidad y que dichos niveles son probablemente exclusivamente determinados por variantes dentro del gen F12. De estas variantes, la más estudiada es la transición 46C>T. Se ha mostrado que hay una asociación entre este polimorfismo y la variación en los niveles del factor XII en el plasma de la sangre, ya que se cree que el 46C>T afecta la eficiencia de la traducción. Asimismo, estudios de caso-control en muestras españolas indicaron que el genotipo T/T de dicho polimorfismo es un factor independiente de riesgo para el desarrollo de trombosis venosa, ataque cerebral isquémico y enfermedad coronaria aguda.

OBJETIVOS

El presente trabajo se diseñó en función de los siguientes objetivos:

- Describir por primera vez la variación que presentan los polimorfismos Alu del cromosoma de X en poblaciones mediterráneas; usar esta variación para investigar las relaciones genéticas interpoblacionales y las posibles influencias subsaharianas, y comprobar la utilidad de estos marcadores para los estudios de genética de poblaciones humanas.
- Analizar la variación molecular dentro y alrededor de la región genómica F7 en grupos humanos de distinta procedencia (Mediterráneo, África subsahariana y Bolivia) con el fin de explorar los patrones de

desequilibrio de ligamiento y sus implicaciones sobre la historia evolutiva de la región promotora del gen F7 en las poblaciones humanas.

- Analizar la variación molecular dentro y alrededor de la región genómica F12 en los mismos grupos humanos y utilizarla conjuntamente con la variación de la región F7 anteriormente mencionada para explorar la estructura genética de las poblaciones humanas en el Mediterráneo.
- Evaluar, como aplicación concreta de la variación analizada de la región F12, el papel del polimorfismo 46C>T en la susceptibilidad a la cardiopatía isquémica en dos muestras del mediterráneo occidental utilizando un doble estudio de asociación genética.

MATERIAL Y METODOS

Poblaciones estudiadas

Los polimorfismos Alu del cromosoma X, se analizaron en seis muestras poblacionales (525 personas): dos grupos de bereberes de Marruecos y Egipto (Alto Atlas y el Oasis de Siwa), uno de Monastir (Túnez), uno del País Vasco y uno de Creta (Grecia). Asimismo, fue incluida una muestra procedente de la Costa de Marfil, como representante de la variación subsahariana.

El trabajo epidemiológico se realizó en 101 familias nucleares de la zona de Barcelona (n = 302) con un hijo afectado por cardiopatía isquémica, además de 76 pacientes de cardiopatía isquémica procedentes de Monastir y 118 individuos sanos procedentes del norte y centro/sur de Túnez.

La variación de las regiones genómicas de los genes F7 y F12 fue determinada en un conjunto de 16 poblaciones (687 indivíduos) de diferentes lugares: las

muestras de España procedían del norte (Asturias, País Vasco, Valles Pasiegos), noreste (Cataluña) y el sur del país (Andalucía). Francia fue representada por una muestra de Toulouse, mientras que Grecia por una muestra de Creta y Turquía por una población urbana de Estambul. En cuanto a África del Norte, las muestras incluyeron tres grupos étnicos bereberes procedentes de Marruecos (Asni y Khenifra del Alto Atlas; Bouhria del Atlas de noroeste), un grupo bereber de Argelia (M'zab) y un grupo urbano de Túnez (Monastir). Asimismo, se tomaron en consideración tres grupos no mediterráneos: uno de la Costa de Marfil y dos poblaciones indígenas distintas lingüísticamente de Bolivia – los aymaras y los quechuas.

El ADN de todos los análisis se extrajo a partir de sangre total. Todos los participantes tenían sus cuatro abuelos nacidos en la misma región. El estudio se realizó de conformidad con las directrices del Comité Ético de la Universidad de Barcelona y con el consentimiento informado de todos los participantes.

Polimorfismos estudiados

Los marcadores polimórficos analizados incluyen:

- 13 polimorfismos Alu del cromosoma X,
- 14 marcadores de la región del gen F7, repartidos en 5 polimorfismos funcionales, situados en el promotor del gen (siendo cuatro de ellos SNPs y un polimorfismo de inserción/deleción de 10 pares de bases) y 9 polimorfismos neutros (6 SNPs y 3 microsatélites) de la región flanqueante, y

- 8 marcadores de la región del gen F12 que incluyen 1 polimorfismo funcional (SNP) situado en el exón 1 del gen F12 y 7 polimorfismos neutros (4 SNPs y 3 microsatélites) de la región flanqueante.

Análisis genotípico

Los polimorfismos Alu se genotiparon mediante una amplificación de PCR seguida por una electroforesis y observación directa de los patrones de migración de las bandas teñidas en gel de agarosa. Todos los SNPs y el polimorfismo de inserción/deleción fueron genotipados con unos ensayos de espectrometría de masa en la plataforma *iPLEX MassARRAY* de Sequenom. En cuanto a los microsatélites, ellos fueron genotipados mediante una amplificación de PCR y un consecutivo análisis de fragmentos en el analizador *Genetic Analyzer* de Applied Biosystems.

Análisis estadístico

En cuanto a los análisis estadísticos:

- Las frecuencias alélicas de todos los polimorfismos se calcularon con GENETIX.
- La variación de los microsatélites (número de alelos, media, varianza y heterocigosidad) se calculó con programa MSA
- El ajuste al equilibrio de Hardy-Weinberg se examinó con ARLEQUIN.
- Las distancias genéticas entre pares de poblaciones en todos los análisis relevantes se calcularon mediante la formula de Reynolds con el programa PHYLIP.
- La fase haplotípica en las regiones genómicas F7 y F12 se infirió con PHASE.

- El desequilibrio de ligamiento entre pares de loci fue evaluada a través de los estadísticos D' y r^2 con Haploview y por la prueba *Black & Krafsur* con GENETIX.
- La corrección de Bonferroni se aplicó en todos los casos de pruebas múltiples.
- Para detectar diferencias significativas entre pares de poblaciones se implementó el programa GENEPOP.
- La estructura intra e interpoblacional se examinó con los valores F_{ST} mediante el análisis de varianza molecular (AMOVA) utilizando ARLEQUIN.
- Las relaciones genéticas entre las muestras estudiadas se visualizaron con un análisis de componentes principales con el paquete estadístico ade4 en R.
- En el estudio de las Alus del cromosoma X se utilizó de manera complementaria al análisis de componentes principales el programa bayesiano STRUCTURE para inferir la estructura poblacional.
- El aislamiento por distancia como posible mecanismo de la diferenciación poblacional observada fue evaluado con una prueba Mantel de correlación entre matrices de distancias genéticas y geográficas utilizando el paquete estadístico ade4 en R (10.000 permutaciones).
- La posibilidad de una barrera genética entre las dos costas del mediterráneo se examinó con un análisis gráfico de la correlación entre distancias geográficas y genéticas.
- El flujo génico subsahariano hacia las dos costas mediterráneas basado en las regiones F7 y F12 se examinó a través (i) del análisis de haplotipos

compartidos entre las distintas poblaciones y (ii) del cálculo de la actividad migratoria entre pares de poblaciones reflejada por el valor Nm según el Modelo de Isla de Fisher.

- Los posibles efectos selectivos fueron investigados con (i) una comparación de valores F_{ST} entre loci neutros y funcionales mediante la prueba no paramétrica de Mann-Whitney con R y (ii) la prueba LRH (Long-range haplotype test) utilizando una aplicación de internet.

RESULTADOS Y CONCLUSIONES

- La variación de las regiones genómicas analizadas indica que la diferenciación genética entre poblaciones del norte de África y sur de Europa es baja pero significativa. Esta observación es más pronunciada en los datos autosómicos (F_{CT} = 0.013, $p<0.05$) que en los datos del cromosoma X (F_{CT} = 0.012, $p>0.05$).
- El análisis gráfico de la correlación entre distancias geográficas y genéticas en el mediterráneo occidental mostró que, para del mismo rango de distancias geográficas, las distancias genéticas entre poblaciones procedentes de lados opuestos son más largas que entre poblaciones de la misma costa. Esta observación sugiere que el mediterráneo puede haber actuado, al menos parcialmente, como una barrera al flujo génico entre el norte de África y sur Europa.
- El análisis de haplotipos de las regiones F7 y F12 ha proporcionado evidencias de un flujo génico subsahariano más elevado hacia el norte de África que el sur de Europa. Esta observación podría ser el resultado de la

existencia de una barrera genética en el Mediterráneo, pero no se puede descartar la posibilidad de un modelo de diferenciación por aislamiento, ya que la correlación entre las matrices de distancias genéticas y geográficas en el norte de África y sur de Europa es significativa.

- Como han mostrado los análisis de componentes principales, la diferenciación genética interpoblacional en la parte europea del mediterráneo es menos pronunciada que en la parte africana; el norte de África presenta un cierto grado de estructura poblacional que es evidente aún cuando el número de marcadores estudiados es bajo.
- A diferencia de estudios anteriores, en los que tradicionalmente algunas poblaciones se consideraron aislados genéticos, la variación genética estudiada en este trabajo, manifestada en todos los análisis de estructura poblacional y para todos los marcadores usados, mostró que los vascos no presentaron ninguna posición genética especial con respecto a otras poblaciones del sur de Europa. Este hallazgo está de acuerdo con otros estudios recientes.
- Según el análisis de componentes principales basado en polimorfismos Alu del cromosoma X, los bereberes egipcios del Oasis de Siwa aparecen como la población más diferenciada dentro del norte de África y con una fuerte influencia subsahariana (según los resultados del análisis del programa STRUCTURE), posiblemente reflejando un importante aislamiento geográfico después de la desertificación del Sahara.
- Todos los datos afirman que, de todas las poblaciones del norte de África, la población de Monastir en Túnez es la más cercana genéticamente a Europa, posiblemente debido a los antecedentes históricos de la región y

a un aislamiento geográfico menos pronunciado en comparación con otros países del norte de África.

- La variación genética del cromosoma X, pero sobretodo de los marcadores autosómicos, mostró que los bereberes procedentes del Magreb (Asni y Khenifra) son notablemente los más diferenciados en el conjunto de las poblaciones norteafricanas examinadas, posiblemente debido a una mayor influencia genética de África subsahariana.
- La distribución poblacional de la variación en la región promotora del gen F7 no presenta rasgos de selección positiva en la región mediterránea. En cambio, hay evidencias sólidas de una la selección positiva en la población indígena de Bolivia que sugieren historias evolutivas diferentes en el seno de las poblaciones humanas actuales.
- A diferencia de estudios previos, la falta de asociación entre el polimorfismo FXII 46C>T y la cardiopatía isquémica en el estadísticamente poderoso estudio caso-control de Túnez sugiere que dicho polimorfismo no es un factor de riesgo universal de la cardiopatía isquémica.

References

A

Abdennaji Guenounou B, Loueslati BY, Buhler S, Hmida S, Ennafaa H, Khodjet-Elkhil H, Moojat N, Dridi A, Boukef K, Ben Ammar Elgaaied A, Sanchez-Mazas A: HLA class II genetic diversity in southern Tunisia and the Mediterranean area. *Int J Immunogenet* 2006, **33**:93-103.

Agrafioti I, Stumpf MP: SNPSTR: a database of compound microsatellite-SNP markers. *Nucleic Acids Res* 2007, **35**:D71-75.

Allender S, Scarborough P, Peto V, Rayner M: European cardiovascular disease statistics. Brussels: European Heart Network; 2008.

Amory S, Dugoujon JM, Despiau S, Roubinet F, El Chenawi F, Blancher A: Identification de trois nouveaux allèles O dans une population berbère de Siwa (Egypte). *Antropo* 2004, **7**:105-112.

Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, Bueno-de-Mesquita HB, Gross M, Helzlsouer K, Jacobs EJ, LaCroix A, Zheng W, Albanes D, Bamlet W, Berg CD, Berrino F, Bingham S, Buring JE, Bracci PM, Canzian F, Clavel-Chapelon F, Clipp S, Cotterchio M, de Andrade M, Duell EJ, Fox JW Jr, Gallinger S, Gaziano JM, Giovannucci EL, Goggins M, González CA, Hallmans G, Hankinson SE, Hassan M, Holly EA, Hunter DJ, Hutchinson A, Jackson R, Jacobs KB, Jenab M, Kaaks R, Klein AP, Kooperberg C, Kurtz RC, Li D, Lynch SM, Mandelson M, McWilliams RR, Mendelsohn JB, Michaud DS, Olson SH, Overvad K, Patel AV, Peeters PH, Rajkovic A, Riboli E, Risch HA, Shu XO, Thomas G, Tobias GS, Trichopoulos D, Van Den Eeden SK, Virtamo J, Wactawski-Wende J, Wolpin BM, Yu H, Yu K, Zeleniuch-Jacquotte A, Chanock SJ, Hartge P, Hoover RN:

Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet* 2009, **41**:986-990.

Angius A, Bebbere D, Petretto E, Falchi M, Forabosco P, Maestrale GB, Casu G, Persico I, Melis PM, Pirastu M: Not all isolates are the same: linkage disequilibrium analysis on Xq13.3 reveals different patterns in Sardinian subpopulations. *Hum Genet* 2002, **111**:9-15.

Arnaiz-Villena A, Gomez-Casado E, Martinez-Laso J. Population genetic relationships between Mediterranean populations determined by HLA allele distribution and a historic perspective. *Tissue Antigens* 2002, **60**:111-121.

Athanasiadis G, Esteban E, Via M, Dugoujon JM, Moschonas N, Chaabani H, Moral P: The X chromosome Alu insertions as a tool for human population genetics: data from European and African human groups. *Eur J Hum Genet* 2007, **15**:578-583.

Athanasiadis G, Esteban G, Gayà Vidal M, Carreras Torres R, Bahri R, Moral P: Polymorphism FXII 46C>T and cardiovascular risk: additional data from Spanish and Tunisian patients. *BMC Res Notes* 2009, **2**:154.

Athanasiadis G, Esteban E, Gayà-Vidal M, Dugoujon JM, Moschonas N, Chaabani H, Bissar-Tadmouri N, Harich N, Stoneking M, Moral P: Different evolutionary histories of the Coagulation Factor VII gene in human populations? *Ann Hum Genet* 2010, **74**:34-45.

Athanasiadis G, González-Pérez E, Esteban E, Dugoujon JM, Stoneking M, Moral P: The Mediterranean Sea as a barrier to gene flow: evidence from variation in and around the F7 and F12 genomic regions. *BMC Evol Biol* 2010, **10**:84.

B

Bach J, Endler G, Winkelmann BR, Boehm BO, Maerz W, Mannhalter C, Hellstern P: Coagulation factor XII activity, activated factor XII, distribution of factor XII C46T gene polymorphism and coronary risk. *J Thromb Haemost* 2008, **6**:291-296.

Bahri R, Esteban E, Moral P, Chaabani H: New insights into the genetic history of Tunisians: Data from Alu insertion and apolipoprotein E gene polymorphisms. *Ann Hum Biol* 2008, **35**:22-33.

Barbujani G, Bertorelle G, Chikhi L: Evidence for Paleolithic and Neolithic gene flow in Europe. *Am J Hum Genet* 1998, **62**:488-492.

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

Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005, **21**:263-265.

Batzer MA, Deininger PL: Alu repeats and human genomic diversity. *Nat Rev Genet* 2002, **3**:370-379.

Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labuda D, Rubin CM, Schmid CW, Zietkiewicz E, Zuckerkandl E: Standardized nomenclature for Alu repeats. *J Mol Evol* 1996, **42**:3-6.

Belkhir K, Borsa P, Goudet J, Chikhi L, Bonhomme F: GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome,

Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France), 1998.

Benson G: Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999, **27**:573-580.

Black WC, Krafsur ES: A FORTRAN program for the calculation and analysis of two-locus linkage disequilibrium coefficients. *Theor Appl Genet* 1985, **70**:491-496.

Bosch E, Calafell F, Perez-Lezaun A, Comas D, Mateu E, Bertranpetit J: Population history of North Africa: Evidence from classical genetic markers. Hu*m Biol* 1997, **69**:295-311.

Bosch E, Calafell F, Pérez-Lezaun A, Clarimón J, Comas D, Mateu E, Martínez-Arias R, Morera B, Brakez Z, Akhayat O, Sefiani A, Hariti G, Cambon-Thomsen A, Bertranpetit J: Genetic structure of north-west Africa revealed by STR analysis. *Eur J Hum Genet* 2000, **8**:360-366.

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

Brandström M, Bagshaw AT, Gemmell NJ, Ellegren H. The relationship between microsatellite polymorphism and recombination hot spots in the human genome. *Mol Biol Evol* 2008, **25**:2579-2587.

C

Calderón R, Pérez-Miranda AM, Fuciarelli M, Scano G, Carrión M, Alfonso-Sánchez MA, Peña JA, Ambrosio B, De Stefano G: Genetic polymorphisms in autochthonous Basques from northern Navarre. *Anthropol Anz* 2006, **64**:173-187.

Callinan PA, Hedges DJ, Salem AH, Xing J, Walker JA, Garber RK, Watkins WS, Bamshad MJ, Jorde LB, Batzer MA: Comprehensive analysis of Alu-associated diversity on the human sex chromosomes. *Gene* 2003, **317**:103-110.

Campuzano V, Montermini L, Moltò MD, Pianese L, Cossée M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Zara F, Cañizares J, Koutnikova H, Bidichandani SI, Gellera C, Brice A, Trouillas P, De Michele G, Filla A, De Frutos R, Palau F, Patel PI, Di Donato S, Mandel JL, Cocozza S, Koenig M, Pandolfo M: Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 1996, **271**:1423-1427.

Cann RL, Stoneking M, Wilson AC: Mitochondrial DnA and human evolution. *Nature* 1987, **325**:31-36.

Capelli C, Redhead N, Romano V, Calì F, Lefranc G, Delague V, Megarbane A, Felice AE, Pascali VL, Neophytou PI, Poulli Z, Novelletto A, Malaspina P, Terrenato L, Berebbi A, Fellous M, Thomas MG, Goldstein DB: Population structure in the Mediterranean basin: a Y chromosome perspective. *Ann Hum Genet* 2006, **70**:207-225.

Cavalli-Sforza LL: The Basque population and ancient migrations in Europe. *Munibe* 1988, **6**:129-137.

Cavalli-Sforza LL, Menozzi P, Piazza A: The History and Geography of Human Genes. Princeton, NJ: Princeton University Press; 1994.

Chen JT, Sokal RR, Ruhlen M: Worldwide analysis of genetic and linguistic relations of human populations. *Hum Biol* 1995, **67**:595-612.

Chen C, Gentles AJ, Jurka J, Karlin S: Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. *Proc Natl Acad Sci USA* 2002, **99**:2930-2935.

Chessel D, Dufour AB, Thioulouse J: The ade4 package-I: one-table methods. *R News* 2004, **4**:5-10.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G: Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. *Proc Natl Acad Sci USA* 1998, **95**:9053-9058.

Comas D, Calafell F, Benchemsi N, Helal A, Lefranc G, Stoneking M, Batzer MA, Bertranpetit J, Sajantila A: Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Hum Genet* 2000, **107**:312-319.

Cooke GS, Hill AV: Genetics of susceptibility to human infectious disease. *Nat Rev Genet* 2001, **2**:967-977.

Crow JF, Kimura M: An introduction to population genetics theory. New York: Harper and Row; 1970.

Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, Dugoujon JM, Crivellaro F, Benincasa T, Pascone R, Moral P, Watson E, Melegh B, Barbujani G, Fuselli S, Vona G, Zagradisnik B, Assum G, Brdicka R, Kozlov AI, Efremov GD, Coppa A, Novelletto A, Scozzari R: Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. *Mol Biol Evol* 2007, **24**:1300-1311.

Cunliffe B: The Ancient Celts. London: Penguin; 1999.

D

de Maat MP, Green F, de Knijff P, Jespersen J, Kluft C: Factor VII polymorphisms in populations with different risks of cardiovascular disease. *Arterioscler Thromb Vasc Biol* 1997, **17**:1918-1923.

Dean AG, Dean JA, Burton AH, Dicker RC: Epi Info™: a general purpose microcomputer program for health information systems. *Am J Preventive Medicine* 1991, **7**:178-182.

Deininger PL., Daniels GR: The recent evolution of mammalian repetitive DNA elements. *Trends Genet* 1986, **2**:76-80.

Deininger PL, Batzer MA: Alu repeats and human disease. *Mol Genet Metab* 1999, **67**:183-193.

Dieringer D, Schlötterer C: Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol Ecol Notes* 2003, **3**:167-169.

Doggen CJ, Rosendaal FR, Meijers JC: Levels of intrinsic coagulation factors and the risk of myocardial infarction among men: opposite and synergistic effects of factors XI and XII. *Blood* 2006, **108**:4045-4051.

E

Edwards MC, Gibbs RA: A human dimorphism resulting from loss of an Alu. *Genomics* 1992, **14**:590-597.

Ellegren H: Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 2004, **5**:435-445.

Endler G, Exner M, Mannhalter C, Meier S, Ruzicka K, Handler S, Panzer S, Wagner O, Quehenberger P: A common C-->T polymorphism at nt 46 in the promoter region of coagulation factor XII is associated with decreased factor XII activity. *Thromb Res* 2001, **101**:255-260.

Ennafaa H, Ben Amor M, Yacoubi-Loueslati B, Khodejt el Khil H, González-Pérez E, Moral P, Maca-Meyer N, Elgaaied A: Alu polymorphisms in Jerba Island population (Tunisia): Comparative study in Arab and Berber groups. *Ann Hum Biol* 2006, **33**:634-640.

Esteban E, Dugoujon JM, Guitard E, Sénégas MT, Manzano C, de la Rúa C, Valveny N, Moral P: Genetic diversity in northern Spain (Basque Country and Cantabria): GM and KM variation related to demographic histories. *Eur J Hum Genet* 1998, **6**:315-324.

Esteban E, González-Pérez E, Harich N, López-Alomar A, Via M, Luna F, Moral P: Genetic relationships among Berbers and South Spaniards based on CD4 microsatellite/Alu haplotypes. *Ann Hum Biol* 2004, **31**:202-212.

Esteban E, Via M, González-Pérez E, Santamaría J, Dugoujon JM, Vona G, Harich N, Luna F, Saetta AA, Bissar N, Moral P: An unexpected wide population variation of the G1733A polymorphism of the androgen receptor gene: data on the Mediterranean region. *Am J Hum Biol* 2005, **17**:690-695.

Esteban E, Rodon N, Via M, Gonzalez-Perez E, Santamaria J, Dugoujon JM, Chennawi FE, Melhaoui M, Cherkaoui M, Vona G, Harich N, Moral P: Androgen receptor CAG and GGC polymorphisms in Mediterraneans: repeat dynamics and population relationships. *J Hum Genet* 2006, **51**:129-136.

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

F

Fadhlaoui-Zid K, Plaza S, Calafell F, Ben Amor M, Comas D, Bennamar El gaaied A. Mitochondrial DNA heterogeneity in Tunisian Berbers. *Ann Hum Genet* 2004, **68**:222-233.

Fair DS: Quantitation of factor VII in the plasma of normal and warfarin-treated individuals by radioimmunoassay. *Blood* 1983, **62**:784-791.

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

Feuk L, Carson AR, Scherer SW: Structural variation in the human genome. *Nat Rev Genet* 2006, **7**:85-97.

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

Foncea N, Gómez Beldarrain M, Ruiz Ojeda J, Carrascosa T, García-Moncó J: Ischemic stroke in a patient with factor XII (Hageman) deficiency. *Neurologia* 2001, **16**:227-228.

G

García-Obregón S, Alfonso-Sánchez MA, Pérez-Miranda AM, Vidales C, Arroyo D, Peña JA: Genetic position of Valencia (Spain) in the Mediterranean basin according to Alu insertions. *Am J Hum Biol* 2006, **18**:187-195.

García-Obregón S, Alfonso-Sánchez MA, Pérez-Miranda AM, de Pancorbo MM, Peña JA: Polymorphic Alu insertions and the genetic structure of Iberian Basques. *J Hum Genet* 2007, **52**:317-327.

Gauderman WJ, Morrison JM: QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies. [http://hydra.usc.edu/gxe], 2006.

Giraldo MP, Esteban E, Aluja MP, Nogués RM, Backés-Duró C, Dugoujon JM, Moral P: Gm and Km alleles in two Spanish Pyrenean populations (Andorra and Pallars Sobira): a review of Gm variation in the Western Mediterranean basin. *Ann Hum Genet* 2001, **65**:537-548.

Girolami A, Randi ML, Gavasso S, Lombardi AM, Spiezia F: The occasional venous thromboses seen in patients with severe (homozygous) FXII deficiency are probably due to associated risk factors: a study of prevalence in 21 patients and review of the literature. *J Thromb Thrombolysis* 2004, **17**:139-143.

González-Pérez E, Via M, Esteban E, López-Alomar A, Mazieres S, Harich N, Kandil M, Dugoujon JM, Moral P: Alu insertions in the Iberian Peninsula and north west Africa - genetic boundaries or melting pot? *Coll Antropol* 2003, **27**:491-500.

González-Pérez E, Esteban E, Via M, Gaya-Vidal M, Athanasiadis G, Dugoujon JM, Luna F, Mesa MS, Fuster V, Kandil M, Harich N, Bissar-Tadmouri N, Saetta A, Moral P: Population relationships in the Mediterranean revealed by autosomal genetic data (Alu and Alu/STR compound systems). *Am J Phys Anthropol* 2009, **141**:430-439.

Guo SW, Thompson EA: Performing the exact test of Hardy-Weinberg proportions for multiple alleles. *Biometrics* 1992, **48**:361-372.

H

Hahn MW, Rockman MV, Soranzo N, Goldstein DB, Wray GA: Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. *Genetics* 2004, **167**:867-877.

Haldane JBS: Disease and evolution. *Ricerca Scientifica* 1949, **19**:3-10.

Harich N, Esteban E, Chafik A, López-Alomar A, Vona G, Moral P. Classical polymorphisms in Berbers from Moyen Atlas (Morocco): genetics, geography,

and historical evidence in the Mediterranean peoples. *Ann Hum Biol* 2002, **29**:473-487.

Harpending H, Cochran G: Genetic diversity and genetic burden in humans. *Infect Genet Evol* 2006, **6**:154-162.

Harpending H, Jenkins T: Genetic distance among southern African populations; in Crawford MH, Workman PL: Methods and Theories of Anthropological Genetics. Albuquerque: University of New Mexico Press; 1973

Harris EE, Hey J: X chromosome evidence for ancient human histories. *Proc Natl Acad Sci USA* 1999, **96**:3320-3324.

Hirszfeld L, Hirszfeld H: Essai d'application des methods au probléme des races. *Anthropologie* 1919, **29**:505-537.

Houck CM, Rinehart FP, Schmid CW: A ubiquitous family of repeated DNA sequences in the human genome. *J Mol Biol* 1979, **132**:289-306.

I

International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005, **437**:1299-1320.

Ishii K, Oguchi S, Murat M, Mitsuyoshi Y, Takeshita E, Ito D, Tanahashi N, Fukuuchi Y, Oosumi K, Matsumoto K, Kitajima M, Yamamoto M, Watanabe G, Ikeda Y, Watanabe K: Activated factor XII levels are dependent on factor XII 46 C/T genotypes and factor XII zymogen levels, and are associated with vascular risk factors in patients and healthy subjects. *Blood Coagul Fibrinol* 2000, **11**:277-284.

Izaabel H, Garchon HJ, Caillat-Zucman S, Beaurain G, Akhayat O, Bach JF, Sanchez-Mazas A: HLA class II DNA polymorphism in a Moroccan population from the Souss, Agadir area. *Tissue Antigens* 1998, **51**:106-110.

J

Jobling M, Hurles M, Tyler-Smith C: Human evolutionary genetics: Origins, peoples, disease. New York: Garland; 2004.

K

Kanaji T, Okamura T, Osaki K, Kuroiwa M, Shimoda K, Hamasaki N, Niho Y: A common genetic polymorphism (46 C to T substitution) in the 5'-untranslated region of the coagulation factor XII gene is associated with low translation efficiency and decrease in plasma factor XII level. *Blood* 1998, **91**:2010-2014.

Kanaji T, Watanabe K, Hattori S, Urata M, Iida H, Kinoshita S, Kayamori Y, Kang D, Hamasaki N: Factor XII gene (F12) -4C/C polymorphism in combination with low protein S activity is associated with deep vein thrombosis. *Thromb Haemost* 2006, **96**:854-855.

Kanaji T: Lower factor XII activity is a 'risk' marker rather than a 'risk' factor for cardiovascular disease: a rebuttal. *J Thromb Haemost* 2008, **6**:1053-1054.

Kandil M, Moral P, Esteban E, Autori L, Mameli GE, Zaoui D, Calo C, Luna F, Vacca L, Vona G: Red cell enzyme polymorphisms in Moroccans and southern Spaniards: new data for the genetic history of the western Mediterranean. *Hum Biol* 1999, **71**:791-802.

Kannel WB, Wolf PA, Garrison RJ: The Framingham Study: an epidemiological investigation of cardiovascular disease. Bethesda: National Heart, Lung and Blood Institute; 1988.

Kayser M, Vowles EJ, Kappei D, Amos W: Microsatellite length differences between humans and chimpanzees at autosomal loci are not found at equivalent haploid Y chromosomal Loci. *Genetics* 2006, **173**:2179-2186.

Keys A, Menotti A, Karvonen MJ, Aravanis C, Blackburn H, Buzina R, Djordjevic BS, Dontas AS, Fidanza F, Keys MH: The diet and the 15-year death rate in the seven countries study. *Am J Epidemiol* 1986, **124**:903-915.

Kim H, Dionne RA: Lack of influence of GTP cyclohydrolase gene (GCH1) variations on pain sensitivity in humans. *Mol Pain* 2007, **3**:6.

Kimura M, Weiss GH: The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* 1964, **49**:561-576.

Kluft C, Dooijewaard G, Emeis JJ: Role of the contact system in fibrinolysis. *Semin Thromb Hemost* 1987, **13**:50-68.

Korenberg JR, Rykowski MC: Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* 1988, **53**:391-400.

Kröpelin S, Verschuren D, Lézine AM, Eggermont H, Cocquyt C, Francus P, Cazet JP, Fagot M, Rumes B, Russell JM, Darius F, Conley DJ, Schuster M, von Suchodoletz H, Engstrom DR: Climate-driven ecosystem succession in the Sahara: the past 6000 years. *Science* 2008, **320**:765-768.

Kruglyak S, Durrett RT, Schug MD, Aquadro CF: Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci USA* 1998, **95**:10774-10778.

L

Laird N, Horvath S, Xu X: Implementing a unified approach to family based tests of association. *Genet Epidemiol* 2000, **19**:36-42.

Landsteiner K: Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Zbl Bakt I Abt* 1900, **27**:357-362.

Latini V, Sole G, Doratiotto S, Poddie D, Memmi M, Varesi L, Vona G, Cao A, Ristaldi MS: Genetic isolates in Corsica (France): linkage disequilibrium extension analysis on the Xq13 region. *Eur J Hum Genet* 2004, **12**:613-619.

Lindqvist PG, Dahlbäck B: Carriership of Factor V Leiden and evolutionary selection advantage. *Curr Med Chem* 2008, **15**:1541-1544.

Lombardi AM, Bortoletto E, Scarparo P, Scapin M, Santarossa L, Girolami A: Genetic study in patients with factor XII deficiency: a report of three new mutations exon 13 (Q501STOP), exon 14 (P547L) and -13C>T promoter region in three compound heterozygotes. *Blood Coagul Fibrinolysis* 2008, **19**:639-643.

Lucotte G, Guerin P, Halle L, Loirat F, Hazout S: Y chromosome DNA polymorphisms in two African populations. *Am J Hum Genet* 1989, **45**:16-20.

M

Mackay J, Mensah G: The Atlas of Heart Disease and Stroke. Geneva: World Health Organization; 2004.

Malécot G: Les Mathématiques de l'hérédité. Paris: Masson; 1948.

Mann HB, Whitney DR: On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947, **18**:50-60.

Mantel N: Detection of disease clustering and a generalized regression approach. *Cancer Res* 1967, **27**:209-220.

Mathias SL, Scott AF, Kazazian HH Jr, Boeke JD, Gabriel A: Reverse transcriptase encoded by a human transposable element. *Science* 1991, **254**:1808-1810.

McNemar Q: Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947, **12**:153-157.

Mellars P: Neanderthals and the modern human colonization of Europe. *Nature* 2004, **432**:461-465.

Moral P, Valveny N, López-Alomar A, Calo C, Kandil M, Harich N, González-Pérez E, Via M, Esteban E, Dugoujon JM, Vona G: Molecular variation at functional genes and the history of human populations--data on candidate genes for cardiovascular risk in the Mediterranean. *Coll Antropol* 2003, **27**:523-536.

Mountain JL, Knight A, Jobin M, Gignoux C, Miller A, Lin AA, Underhill PA: SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference

of population history and mutational processes. *Genome Res* 2002, **12**:1766-1772.

Mourant AE, Kopec AC, Domaniewska-Sobczak K: The Distribution of the Human Blood Groups and Other Polymorphisms. Oxford: Oxford University Press; 1976.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: A fine-scale map of recombination rates and hotspots across the human genome. *Science* 2005, **310**:321-324.

Myres NM, Ekins JE, Lin AA, Cavalli-Sforza LL, Woodward SR, Underhill PA: Y-chromosome short tandem repeat DYS458.2 non-consensus alleles occur independently in both binary haplogroups J1-M267 and R1b3-M405. *Croat Med J* 2007, **48**:450-459.

N

Norwich JJ: The Middle Sea: A History of the Mediterranean. London: Vintage; 2007.

O

Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, Scozzari R, Cruciani F, Behar DM, Dugoujon JM, Coudray C, Santachiara-Benerecetti AS, Semino O, Bandelt HJ, Torroni A: The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science* 2006, **314**:1767-1770.

P

Peña JA, Garcia-Obregon S, Perez-Miranda AM, De Pancorbo MM, Alfonso-Sanchez MA: Gene flow in the Iberian Peninsula determined from Y-chromosome STR loci. *Am J Hum Biol* 2006, **18**:532-539.

Pereira RW, Santos SS, Pena SD: A novel polymorphic Alu insertion embedded in a LINE 1 retrotransposon in the human X chromosome DXS225: identification and worldwide population study. *Genet Mol Res* 2006; **5**:63-71.

Pérez-Miranda AM, Alfonso-Sánchez MA, Kalantar A, García-Obregón S, de Pancorbo MM, Peña JA, Herrera RJ: Microsatellite data support subpopulation structuring among Basques. *J Hum Genet* 2005, **50**:403-414.

Picornell A, Gómez-Barbeito L, Tomàs C, Castro JA, Ramon MM: Mitochondrial DNA HVRI variation in Balearic populations. *Am J Phys Anthropol* 2005, **128**:119-130.

Plaza S, Calafell F, Helal A, Bouzerna N, Lefranc G, Bertranpetit J, Comas D: Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann Hum Genet* 2003, **67**:312-328.

Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000, **155**:945-959.

Q

Quintana-Murci L, Veitia R, Fellous M, Semino O, Poloni ES: Genetic structure of Mediterranean populations revealed by Y-chromosome haplotypes analysis. *Am J Phys Anthropol* 2003, **121**:157-171.

R

Raymond M, Rousset F: GENEPOP version 1.2 population genetics software for exact tests and ecumenicism. *J Hered* 1995; **86**:248-249.

Renfrew C: Before Babel, speculations on the origins of linguistic diversity. *Cambridge Archaeol J* 1991, **1**:3-23.

Reynolds J, Weir BS, Cockerman CC: Estimation of the coancestry coefficient: basis for a short term genetic distance. *Genetics* 1983, **105**:767-779.

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, Beckman G, Beckman L, Bertranpetit J, Bosch E, Bradley DG, Brede G, Cooper G, Côrte-Real HB, de Knijff P, Decorte R, Dubrova YE, Evgrafov O, Gilissen A, Glisic S, Gölge M, Hill EW, Jeziorowska A, Kalaydjieva L, Kayser M, Kivisild T, Kravchenko SA, Krumina A, Kucinskas V, Lavinha J, Livshits LA, Malaspina P, Maria S, McElreavey K, Meitinger TA, Mikelsaar AV, Mitchell RJ, Nafa K, Nicholson J, Nørby S, Pandya A, Parik J, Patsalis PC, Pereira L, Peterlin B, Pielberg G, Prata MJ, Previderé C, Roewer L, Rootsi S, Rubinsztein DC, Saillard J, Santos FR, Stefanescu G, Sykes BC, Tolun A, Villems R, Tyler-Smith C, Jobling MA: Y-chromosomal diversity in Europe is clinical and

influenced primarily by geography, rather than by language. *Am J Hum Genet* 2000, **67**:1526-1543.

Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, Batzer MA, Deininger PL: Alu insertion polymorphisms for the study of human genomic diversity. *Genetics* 2001, **159**:279-290.

S

Sabater-Lleal M, Almasy L, Martínez-Marchán E, Martínez-Sánchez E, Souto R, Blangero J, Souto J, Fontcuberta J, Soria JM: Genetic architecture of the F7 gene in a Spanish population: implication for mapping complex diseases and for functional assays. *Clin Genet* 2006, **69**:420-428.

Sabater-Lleal M, Chillón M, Howard TE, Gil E, Almasy L, Blangero J, Fontcuberta J, Soria JM: Functional analysis of the genetic variability in the F7 gene promoter. *Atherosclerosis* 2007, **195**:262-268.

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: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002, **419**:832-837.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N: Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 1985, **230**:1350-1354.

Sanchez-Velasco P, Gomez-Casado E, Martinez-Laso J, Moscoso J, Zamora J, Lowy E, Silvera C, Cemborain A, Leyva-Cobián F, Arnaiz-Villena A: HLA alleles in

isolated populations from North Spain: origin of the Basques and the ancient Iberians. *Tissue Antigens* 2003, **61**:384-392.

Santamaría A, Martínez-Rubio A, Mateo J, Tirado I, Soria JM, Fontcuberta J: Homozygosity of the T allele of the 46 C-->T polymorphism in the F12 gene is a risk factor for acute coronary artery disease in the Spanish population. *Haematologica* 2004, **89**:878-879.

Santamaría A, Mateo J, Tirado I, Oliver A, Belvís R, Martí-Fábregas J, Felices R, Soria JM, Souto JC, Fontcuberta J: Homozygosity of the T allele of the 46 C->T polymorphism in the F12 gene is a risk factor for ischemic stroke in the Spanish population. *Stroke* 2004, **35**:1795-1799.

Schaffner SF: The X chromosome in population genetics. *Nat Rev Genet* 2004, **5**:43-51.

Schlötterer C. The evolution of molecular markers--just a matter of fashion? *Nat Rev Genet* 2004, **5**:63-69.

Schlotterer C: Hitchhiking mapping — functional genomics from the population genetics perspective. *Trends Genet* 2003, **19**:32-38.

Schneider S, Roessli D, Excoffier L: Arlequin: A software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, Geneva, 2000.

Scozzari R, Cruciani F, Pangrazio A, Santolamazza P, Vona G, Moral P, Latini V, Varesi L, Memmi MM, Romano V, De Leo G, Gennarelli M, Jaruzelska J, Villems R, Parik J, Macaulay V, Torroni A: Human Y-chromosome variation in the western

Mediterranean area: implications for the peopling of the region. *Hum Immunol* 2001, **62**:871-884.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA: The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 2000, **290**:1155-1159.

Simoni L, Gueresi P, Pettener D, Barbujani G: Patterns of gene flow inferred from genetic distances in the Mediterranean region. *Hum Biol* 1999, **71**:399-415.

Smit AF: The origin of interspersed repeats in the human genome. *Curr Opin Genet* 1996, **6**:743-748.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, Torroni A, Richards MB: The archaeogenetics of Europe. *Curr Biol* 2010, **20**:R174-183.

Sokal RR, Rohlf FJ: Biometry. New York: Freeman and Co; 1981.

Soria JM, Almasy L, Souto JC, Bacq D, Buil A, Faure A, Martínez-Marchán E, Mateo J, Borrell M, Stone W, Lathrop M, Fontcuberta J, Blangero J: A quantitative trait locus in human factor XII gene influences both plasma factor XII levels and susceptibility to thrombotic disease. *Am J Hum Genet* 2002, **70**:567-574.

Soria JM, Almasy L, Souto JC, Sabater-Lleal M, Fontcuberta J, Blangero J: The F7 gene and clotting factor VII levels: dissection of a human quantitative trait locus. *Hum Biol* 2005, **77**:561-575.

Soria JM, Fontcuberta J: New approaches and future prospects for evaluating genetic risk of thrombosis. *Haematologica* 2005, **90**:1212-1222.

Souto JC, Almasy L, Borrell M, Garí M, Martínez E, Mateo J, Stone WH, Blangero J, Fontcuberta J: Genetic determinants of hemostasis phenotypes in Spanish families. *Circulation* 2000, **101**:1546-1551.

Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Gen* 1993, **52**:506-516.

Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001, **68**:978-989.

Stephens M, Donnelly P: A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003, **73**:1162-1169.

Sukarova E, Dimovski AJ, Tchacarova P, Petkov GH, Efremov GD: An Alu insert as the cause of a severe form of hemophilia A. *Acta Haematol* 2001, **106**:126-129.

T

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM: Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 2007, **17**:520-526.

Teugels E, De Brakeleer S, Goelen G, Lissens W, Sermijn E, De Grève J: De novo Alu element insertions targeted to a sequence common to the BRCA1 and BRCA2 genes. *Hum Mutat* 2005, **26**:284.

Tills D, Kopec AC, Tills R: The distribution of the human blood groups and other polymorphisms. Oxford: Oxford University Press; 1983.

Tirado I, Soria JM, Mateo J, Oliver A, Souto JC, Santamaria A, Felices R, Borrell M, Fontcuberta J: Association after linkage analysis indicates that homozygosity for the 46C-->T polymorphism in the F12 gene is a genetic risk factor for venous thrombosis. *Thromb Haemost* 2004, **92**:892-893.

Tomas C, Sanchez JJ, Barbaro A, Brandt-Casadevall C, Hernandez A, Ben Dhiab M, Ramon M, Morling N: X-chromosome SNP analyses in 11 human Mediterranean populations show a high overall genetic homogeneity except in North-west Africans (Moroccans). *BMC Evol Biol* 2008, **8**:75.

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, Villems R, Kivisild T, Metspalu E, Parik J, Tolk HV, Tambets K, Forster P, Karger B, Francalacci P, Rudan P, Janicijevic B, Rickards O, Savontaus ML, Huoponen K, Laitinen V, Koivumäki S, Sykes B, Hickey E, Novelletto A, Moral P, Sellitto D, Coppa A, Al-Zaheri N, Santachiara-Benerecetti AS, Semino O, Scozzari R: A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet* 2001, **69**:844-852.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ: Harvesting the fruit of the human mtDNA tree. *Trends Genet* 2006, **22**:339-345.

Tunstall-Pedoe H: MONICA Monograph and Multimedia Sourcebook. Geneva: World Health Organization; 2003.

U

Ullu E, Tschudi C: Alu sequences are processed 7SL RNA genes. *Nature* 1984, **312**:171-172.

V

Via M, González-Pérez E, Esteban E, López-Alomar A, Vacca L, Vona G, Dugoujon JM, Harich N, Moral P: Molecular variation in endothelial nitric oxide synthase gene (eNOS) in western Mediterranean populations. *Coll Antropol* 2003, **27**:117-124.

Via M, López-Alomar A, Valveny N, González-Pérez E, Bao M, Esteban E, Pintó X, Domingo E, Moral P: Lack of association between eNOS gene polymorphisms and ischemic heart disease in the Spanish population. *Am J Med Genet A* 2003, **116**:243-248.

Via M, Valveny N, López-Alomar A, Athanasiadis G, Pintó X, Domingo E, Esteban E, González-Pérez E, Moral P: E65 K polymorphism in KCNMB1 gene is not associated with ischaemic heart disease in Spanish patients. *J Hum Genet* 2005, **50**:604-606.

W

Wain LV, Armour JA, Tobin MD: Genomic copy number variation, human health, and disease. *Lancet* 2009; **374**:340-350.

Weatherall DJ: Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. *Nature Rev Genet* 2001, **2**:245-255.

Weiss KM: Tilting at quixotic trait loci (QTL): an evolutionary perspective on genetic causation. *Genetics* 2008, **179**:1741-1756.

Weiss KM, Long JC: Non-Darwinian estimation: my ancestors, my genes' ancestors. *Genome Res* 2009, **19**:703-710.

Wilder J, Hollocher H: Mobile elements and the genesis of microsatellites in dipterans. *Mol Biol Evol* 2001, **18**:384-392.

Wright S: Size of population and breeding structure in relation to evolution. *Science* 1938, **87**:430-431.

Wright S: Isolation by distance. Genetics 1943, **28**:114-138.

Wright S: The genetical structure of populations. *Ann Eugen* 1951, **15**:323-354.

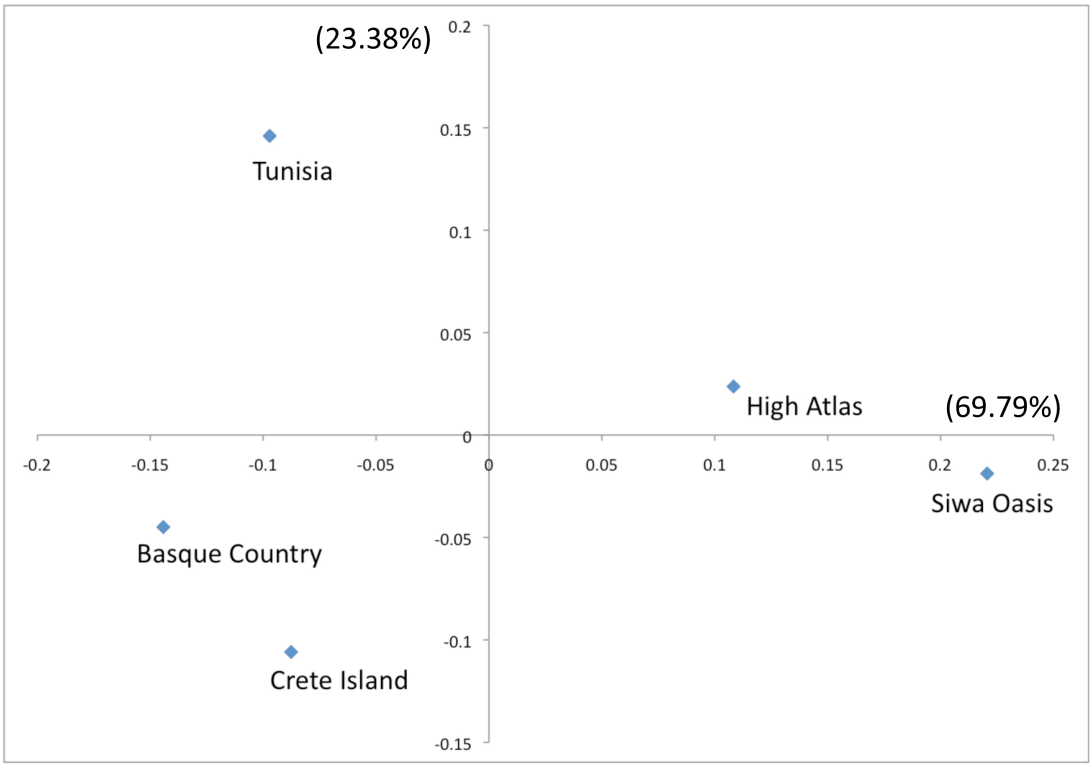
Wright S: Evolution and the Genetics of Populations II, The Theory of Gene Frequencies. Chicago: University of Chicago Press; 1969.

Y

Yu N, Fu YX, Li WH: DNA polymorphism in a worldwide sample of human X chromosomes. *Mol Biol Evol* 2002; **19**:2131-2141.

Appendix

Appendix 1: Additional file to the Athanasiadis et al., 2007 manuscript



Principal components analysis of the five Mediterranean groups (Basque Country, Crete Island, High Atlas, Siwa Oasis and Tunisia) based on the variation from the 13 X chromosome Alu insertions (Athanasiadis et al., 2007). The first two axes account for 93.17% of the total variance. The percentage of variance accounted for by each axis is shown in brackets.

Appendix 2: Links to the additional files provided with the Athanasiadis et al., 2010b manuscript

- Additional file 1: Additional file 1.doc, 70K
  http://www.biomedcentral.com/imedia/9392670163694702/supp1.doc
- Additional file 2: Additional file 2.doc, 93K
  http://www.biomedcentral.com/imedia/8344071033694703/supp2.doc
- Additional file 3: Additional file 3.doc, 64K
  http://www.biomedcentral.com/imedia/1157047117369471/supp3.doc
- Additional file 4: Additional file 4.xls, 99K
  http://www.biomedcentral.com/imedia/4472423143694724/supp4.xls
- Additional file 5: Additional file 5.xls, 37K
  http://www.biomedcentral.com/imedia/1587556731369471/supp5.xls
- Additional file 6: Additional file 6.doc, 49K
- http://www.biomedcentral.com/imedia/1059130297369472/supp6.doc

Appendix 3: Additional file to the Athanasiadis et al., 2007 manuscript

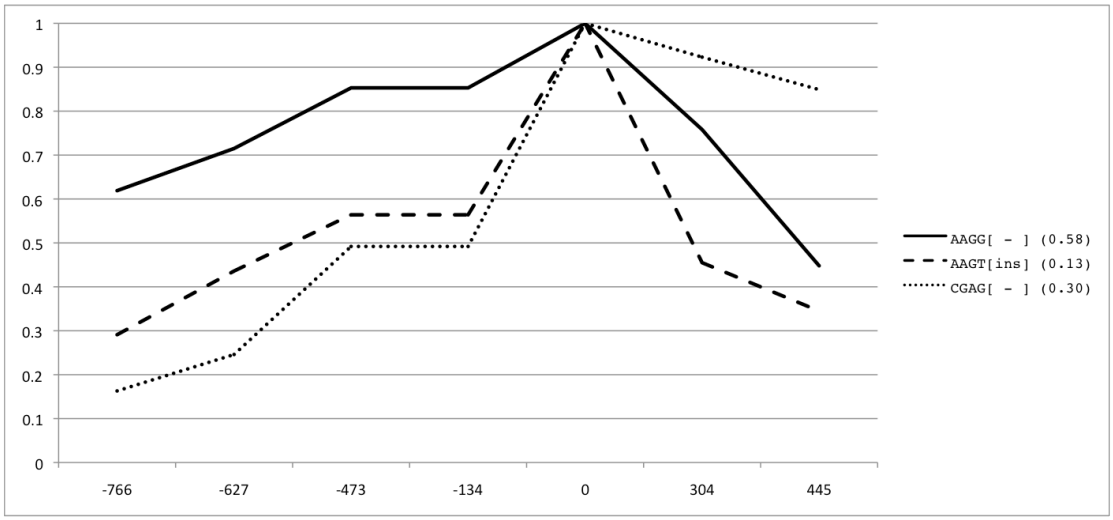


Figure AP2: The extended haplotype homozygosity pattern at varying distances from the core region, for each of the three most frequent core haplotypes from the F7 promoter region for the Siberian Yakuts (Athanasiadis et al., 2010a). The value in brackets indicates haplotype frequency.

Appendix 4: Errata in the published articles

-Athanasiadis et al., 2007:

- **Discussion, second paragraph**: '(FCT values: 1.80 and 1.96%, respectively)' should be changed to '(FCT values: 1.96 and 1.80%, respectively)'.
- **Thoughout the text**: References to 'Calliman et al.' should be changed to 'Callinan et al.'.

-Athanasiadis et al., 2009:

- **Findings, first paragraph**: 'Coagulation factor XVII' should be changed to 'Coagulation factor XII'.
- **Findings, fifth paragraph**: 'North and Central-North' should be changed to 'North and South-Central'.

-Athanasiadis et al., 2010a:

- **Figure 1 legend**: 'MZ: M'zab, Algeria (41)' should be changed to 'MZ: M'zab, Algeria (31)'.
- **Figure 1 legend**: 'TN: Monastir, Tunisia (42)' should be changed to 'TN: Monastir, Tunisia (41)'.