UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

FACULTADE DE MEDICINA E ODONTOLOXÍA

INSTITUTO DE CIENCIAS FORENSES "LUÍS CONCHEIRO"

Genetic analysis of uniparental and autosomal markers in human populations

Memoria que presenta, para optar al grado de doctor,

Francesca Brisighelli

Santiago de Compostela, 2012

El Doctor D. Ángel Carracedo Álvarez, Catedrático de Medicina Legal de la Facultad de Medicina de la Universidad de Santiago de Compostela, y el Doctor D. Antonio Salas Ellacuriaga, Profesor Titular en la Universidad de Santiago de Compostela,

CERTIFICAN:

Que la presente memoria que lleva por título **"Genetic analysis of uniparental and autosomal markers in human populations",** de la licenciada en Sciencias Naturales por la Universidad de Roma "La Sapienza" *Francesca Brisighelli*, ha sido realizada bajo nuestra dirección, considerándola en condiciones para optar al Grado de Doctor y autorizándola para su presentación ante el Tribunal correspondiente. Y para que así conste, firmamos la presente en Santiago de Compostela, .

Prof. Dr. Ángel Carracedo Álvarez

Dr. Antonio Salas Ellacuriaga

Francesca Brisighelli

ai miei genitori

Ringraziamenti

….e' proprio vero, solo adesso che mi ritrovo di fronte questo foglio bianco, mi rendo conto che scrivere questa parte non è affatto facile! Spero di non dimenticare nessuna delle persone che mi hanno accompagnato in questo meraviglioso percorso, non solo dal punto di vista lavorativo, ma anche, e soprattutto, di crescita personale.

Iniziamo…

Grazie,

a Ángel, per aver creduto in me, dandomi la possibilità di unirmi al tuo gruppo di lavoro e poter così anche sperimentare ciò che significa vivere fuori dal proprio paese

a Vincenzo, per avermi, in quel momento, "lasciato andare" dal laboratorio di Roma, incitandomi ad iniziare questa impegnativa ma incredibile esperienza

a Maviky, per l'appoggio incondizionato, per le chiacchiere e per i consigli, per ciò che riguarda il lavoro ma anche il resto

a Toño, "el Xefiño": mi risulta difficile riassumere in 2 righe la gratitudine che ti devo. La tua dedizione alla ricerca mi è stata d'esempio in tutti questi anni. Grazie per l'aiuto, per il sostegno e per la molta pazienza dimostrata con una studentessa italiana

un po' irriverente. Senza di te gran parte di questo lavoro non sarebbe stato possibile

a Cristian, per aver sempre riposto in me estrema fiducia, per la disponibilità dimostrata e per i consigli dati nei momenti in cui dovevo prendere decisioni importanti. Anche nel tuo caso, la passione incondizionata che nutri per la ricerca e lo studio mi hanno sempre spronato a dare il meglio. Grazie per l'esperienza nel laboratorio di Oxford: sono stati due anni "tosti", ma allo stesso tempo formativi. Naturalmente senza di te, "l'altra" parte di questo lavoro non avrebbe modo di esistere

a Giovanni, per avermi fatto appassionare al lavoro di laboratorio, per la fiducia, per gli scambi di idee, a volte un po' accesi, ma sempre produttivi, per l'onestà ed equità dimostrata in tutte le situazioni.

al laboratorio di Santiago:

a quelli che c'erano quando sono arrivata: Chris, Maria B, Bea S., Gloria, Alex, Nuria, Eva (come ci manchi qui su).

a tutti quelli arrivati con me e successivamente: Manuel, Hannah, Cata, Alberto, Alejandro, Yari, LuisMa, Ana F., Olalla, Jens, Miguel, Ceres (la Fantastish) Montse (grazie per avermi accolto in casa quando non sapevo dove sbattere la testa), Rocio, Maria D., Angela (last but not least).

…e adesso in ordine sparso:

a Paula, che dire, per la "fratellanza", per le mille avventure, chiacchiere, risate, consigli, per l'aiuto, per la musica, per tutto. La mia esperienza santiaghesa non sarebbe stata la stessa senza di te

a Raquel, mi Raqui, non ho parole per ringraziarti e per esprimere tutto l'affetto che nutro nei tuoi confronti: sei una delle persone più dolci e sensibili che conosca, e per questo avrai sempre un posto importante nel mio cuore

a Meluccia, per essere così come sei: simply special!!

a Maria C., per l'appoggio, per aver da subito capito il mio "itagnolo", per le mille conversazioni, per le partite di football, per l'esperienza ad Oxford, per le caramelle verdi, ma più semplicemente per l'affetto reciproco incondizionato

a Vane e Anita, per l'aiuto, per la pazienza, per l'amicizia sincera, per apprezzare la mia "bruschetta", e, soprattutto, per la sopportazione: comunque per il "por" e "para" non c'è soluzione, lo siento

a Fonde, perché alla fine siamo 2 frikis. Grazie per tutte le conversazioni assurde durante i nostri pranzi da Rudy…per la scienza e la fantascienza, e per la scrivania che mi hai permesso occupare negli ultimi mesi

ai miei due "figliocci", Carla e Danel, per le risate, i pranzi in biblioteca, le conversazioni iperboliche e molto,molto di più. Vi auguro possiate realizzare tutto ciò che più desiderate, siete 2 cracks

a Ulises e Iva, e al "trio dei granchietti", perché anche se lontani geograficamente, un filo ci unirà sempre

agli amici di Santiago:

a Pablo e Elena, grazie per tutta l'allegria e l'energia positiva nei momenti di "decompressione" tra il lavoro e casa. Siete unici!

a Claudia e Cristian, a Ingrid, a Thor e al mitico "Padrino", a Isabel, alle "tossiche" Ana e Marta un grazie speciale per avermi accolto nelle vostre case, a Paco, a Tere

al laboratorio dell'Università Cattolica di Roma, soprattutto a Ilaria, Francesca, Laura e Federica

al laboratorio della Sapienza di Roma: a Valentina, Chiara Paolo e Cinzia

al laboratorio di Oxford: a George e a Sarah, per la reciproca pazienza

agli amici di Oxford, o meglio alla mia famiglia di Oxford, in particolare a Estrella e Israel, a Juan, a Ana e Vito: grazie per avermi fatto sentire a casa

alle mie amiche-sorelle di Roma: Francesca, Lisa (e le altre Relle), Tiziana, Laura, Maria Laura, Giorgia, Donatella, Antonella, Marina, Raffaella, Valeria per sopportare le mie apparizioni e sparizioni, le mie presenze e assenze, naturalmente solo fisiche

a Franca, a tutti i tuoi consigli ma soprattutto alle mille risate

a Elita, per l'aiuto e per esserci sempre nei momenti complicati

a mio fratello Paolo, per essere sempre il mio specchio fedele

a mio padre e a mia madre, per avermi sempre dato la possibilità e la libertà di scegliere ciò che più avrei voluto fare nella vita, per non avermi mai condizionato e per esservi rassegnati ad avere una figlia nomade: per tutto questo mi considero una privilegiata. Non sarei la stessa persona senza i vostri insegnamenti al rispetto, all' onestà e alla lealtà.

Table of contents

Abbreviations

AB	Applied Biosystem
AIM	Ancestry Informative Markers
aDNA	Ancient DNA
Bp	Base pair
DNA	Deoxyribonucleic Acid
FP	Food Producers
HG	Hunter-gathers
HLA	Human Leukocyte Antigens
HVR	Hypervariable Region
HVS	Hypervariable Segment
INDEL	Insertion-Deletion polymorphism
ISFG	International Society for Forensic Genetics
Kya	Thousands years ago
Kb	Kilo bases
LGM	Last Glacial Maximum
LINE	Long Nuclear Interspersed elements
LD	Linkage Disequilibrium
LR	Likelihood Ratio
Mb	Mega bases
mtDNA	mitochondrial DNA
MSY	Male Specific region
Ne	Effective Population size
NIST	National Institute of Standards and Technology
NRPY (NRY)	Non-Recombining Portion of the Y chromosome
PAR	Pseudoautosomal Region
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
SINE	Short Nuclear Interspersed elements
SMM	Step-wise Mutation Model
SNP	Single Nucleotide Polymorphism
SRY	Sex-determining Region of the Y chromosome
SSR	Simple Sequence Repeats
STR	Short Tandem Repeat
SWGDAM	Scientific Working Group on DNA Analysis Methods
TMRCA	Time More Recent Common Ancestor
UEP	Unique Polymorphisms Event
Ybp	Years before present
YCC	Y Cromosome Consortium
YHRD	Y Chromosome Haplotype Reference Database
Y-STR	Short Tandem Repeat of the Y chromosome
VNTR	Variable Number Tandem Repeat

Introduction

1. **Introduction**

Population genetic aims at understanding the forces that determine evolution, trying to reconstruct the history of human populations. The characterization of the distribution of genetic variability, within different populations, in the various regions of the globe, allow to investigate the nature of genetic affinities, or even of phylogenetical proximity, within and among the populations themselves. The most useful information comes from "genetic polymorphisms" - variants transmitted in strict mendelian fashion - but in order to understand the history and evolution of populations it is usually necessary to study a large number of them.

The study of how genetic variability distributes among and within populations is a key aspect for association studies in disease studies, as well as for forensic genetic purposes. In this latter case, the study of genetic variability is essential for individual identification through a genetic 'fingerprint', i.e. a set of markers can be so variable that the observed allelic combination are nearly individual specific. Therefore, the advent of the genomic era has shifted the interest of the genetic scientific community from the proteins to the polymorphisms of sequence as tools of investigation. There are a few hundred of studies around worldwide human populations on autosomal, mtDNA and Y chromosome data. The number of polymorphisms discovered in the last years is enormous, also as a consequence of the development of the Human Genome Projet (HGP), an international collaborative project launched in the 1990s, aiming to describe the whole human genomic variablity. The HGP ultimate goals were to identify all the more than 20,000 human genes and render them accessible for further biological studies, determine the sequences of the 3 billion chemical base pairs that make up human DNA and store all these informations in databases. Though the HGP was completed in 2003, analyses of the data will continue for many years. A particular effort has recently been devoted also to the development of new fast, reliable and, as far as possible, automated techniques, that could speed up SNPs genotyping.

1.1 Population genetics and forensic genetics

1.1.1 Aims

Starting from the second half of the 20th century, anthropologists realised the possibility of dealing with diversities between individuals studying the dynamic processes that are at the base of humanity differentiation. From a descriptive approach, the transformation to a perciving setting due to the biological sciences development, put the bases for the Anthropological evolution origin. The research methods begin to make use of new possibilities thanks to the introduction of the genetic polimorphisms. The biomolecular approach offers the possibility to analyze monogenic systems whose phenotipic expression does not suffer of environmental infuences, and consequently demostrating a direct correlation between genotype and phenotype. Genetic polimorphisms are determined from the genetic variability of biochemical components, like seric and red cells proteins and antigene membrane, and are generally under control of a single gene. Excluding some special cases, the majority of the genetic polimorphisms is subjected to neutral evolution. For that reason, nowadays they are considered very useful in order to investigate the genetic proximity of different populations and in order to understand the role of microevolutionary factors on neutral basis. Population genetics is the scientific enquiry of this central problem in biology and it studies the distribution of genetic diversity among populations. The principal population genetics aims are: 1) describing the distribution of the genetic diversity beard by modern populations and of its apportionment among subpopulations, i.e. population structure; 2) inferring the pre-historical and historical events that determined the observed modern diversity and structure. The inference process, based on population genetics' models, is strongly motivated by the anthropological interest in the history of our species, its origins, movements and demographic development.

Also forensic genetics, the science combining population genetics and forensic medicine, is using the genetic variability of humans. In this case we have two main applications: 1) the individual identification for criminal cases and 2) the identification of two individuals as close relatives, such as parent and child, through a genetic 'fingerprinting'. The term 'fingerprinting' of an individual was first introduced by Jeffreys et al. (1985) to designate a sequence of allelic states, at the analyzed loci, that are nearly unique to a given individual. In some cases, the genetic makeup of an individual cannot be considered in

isolation, but has to be related to the degree and structure of genetic variation present in the population to which that individual belongs. The employed methods of DNA typing, in fact, cannot guarantee that the given genotype is unique and that there is no other person carrying the same markers. Because of this, probabilities are computed: e.g., the probability that a person has left a biological sample in the criminal scene, or the probability that the presumed father is the biological father of a child. Estimating the probabilities is based on knowing the genotype frequencies of the population, to which the people involved in the case, belong. The interpopulation differences in allele or genotype frequencies, the genetic disequilibrium in the populations, the relatedness of a persons involved in the case, and other factors are required in population genetic analysis, since disregarding them may significantly affect the probability value.

1.1.2 Common aspects and tools of investigation

The introduction of the DNA polymorphisms in forensic work constitutes one of the great legal medicine revolutions, because it permits to strictly establish the kinship relation and identitification between two individuals.

The term polymorphism was first described by Ford in 1940 as "the combined appearance in a place of two or more discontinuous forms of the same species, in such a way that the rarest between them cannot be maintained simply through a periodic mutation". In general terms, a locus is considered polymorphic when the most common allele for this locus has a frequency inferior to 99%. Generally, an allele is an alternative form of a gene (one member of a pair) that is located at a specific position on a specific chromosome. These DNA codings determine distinct traits that can be passed on from parents to offspring. The process by which alleles are transmitted was discovered by Gregor Mendel and formulated in what is known as Mendel's law of segregation.

The classic polymorphisms were the first genetic variability markers used in the forensic field, like for example the AB0 system discovered by Landsteiner in 1900. In 1980 Wyman and White introduced the concept of genetic identification through RFLPs. The use of RFLPs of VNTR loci, replaced the use of the classic protein markers. The DNA presented a great chemical stability in comparison with the fast degradation of the classic markers, and its high variability immediately turned it in the favourite working tool. Shortly after the

discovery of the "DNA fingerprint", the PCR introduction (Mullis and Faloona 1987; Saiki et al. 1988) allowed the hypervariable loci, minisatellites and micro-satellites (STRs) analysis from biological remains, that could not be analyzed until that moment, because of the limited amount of DNA that could be obtained from the majority of the forensic scenarios.

The frequency of the polymorphic variants in a population is regulated by a series of factors, that are: the natural selection and some stochastic factors, like genetic drift, gene flow and mutation. The genome has a division related to the evolutionary processes regulating the variability of each region: this division is a direct reflection of the functions of the differente parts. In that way, the gene coding regions and the regions with regulating functions of genic expression are in the first group, and their variability is strongly influenced by natural selection; on the other hand, there are regions with a not well-known function, commonly known as "junk" DNA, that are not selective, whose variability is only regulated by stochastic processes. Several mathematical models explain the way the evolutionary forces enact on the genome and which of them determine changes. Nowadays, Motô Kimura theories are accepted (The neutral theory of molecular evolution, (1983)), according to which the evolution of most of the genome would be neutral, regulated only by stochastic processes, in a small part only influenced by a negative natural selection process.

The genome selective regions are usually less polymorphic than the non-selective ones, due to the processes regulating their evolution (Akey et al. 2004; Bowcock et al. 1991). Part of the variants appearing in the functional regions can even bring an error in the resulting protein expression sequence or in the gene regulation, leading to the drastic reduction of the individual "fitness" or even its premature death: for that reason this variant would be eliminated or its frequency reduced by negative natural selection. On the other hand, the change could confer a clear advantage in the survival or reproduction possibilities of the individuals carring it, reaching a higher frequency in the following generations through a positive natural selection process. Anyway, only a small part of the changes occurred in these zones will not affect the individuals "fitness" and could be maintained as polymorphisms.

On the contrary, in the non-selective zones, for the fact they do not have any function, almost any "new" variant could merge to the population gene pool, and at the same time, all the possible variants would tend to present frequency values relatively low and

stable. The evolutionary forces acting in these zones are: migration, genetic drift and mutation.

Migration uniforms the frequencies, introducing previously non-existenting changes in the receiving population gene pool, or increasing their frequency. The drift and the mutation, acting randomly, can cause changes in the gene frequencies either increasing the differences between populations or reducing them. Genetic drift or allelic drift is the change in the frequency of a gene variant (allele) in a population, due to random sampling. The alleles in the offspring are a sample of those in the parents, and chance has a role in determining whether a given individual survives and reproduces. Genetic drift may cause gene variants to disappear completely, and thereby reduce genetic variation. Generally the genetic drift will act increasing the variance between groups and diminishing the intragroup one. Mutation occurs with a certain probability in each different nucleotide position, although this probability is usually very low, so that it is probable that a change occurs in a population and not in others, increasing the interpopulation variation. Generally, the mutation and the migration introduce changes in the gene pool of a population (the first increasing the heterogeneity and the second the interpopulation homogeneity) and the drift modifies the frequencies randomly, fixing or eliminating the introduced changes.

The DNA polymorphisms used in forensic applications, depend on the non-selective zones of the genome characteristics (Santos and Tyler-Smith 1996).

Most of the genome is biparentally inherited and recombines. However, two particular segments of the DNA are inherited from one parent only and do not recombine: the mtDNA and, for the most of its length, the Y chromosome. Figure 1.1 shows the different patterns of inheritance for the mtDNA, Y chromosome and autosomes.

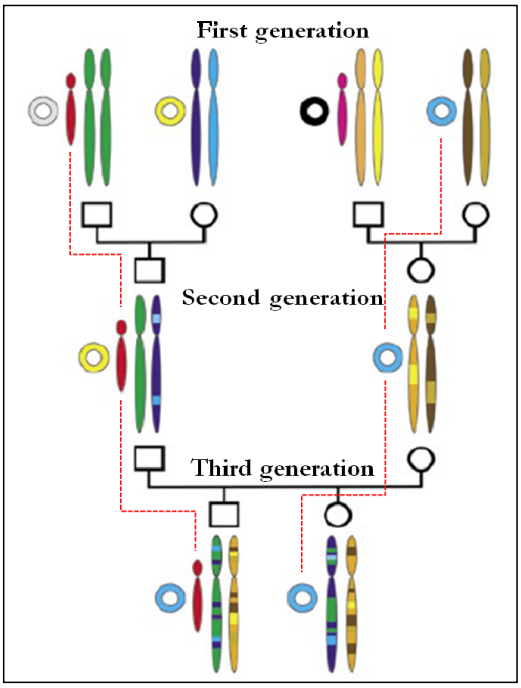


Figure 1.1 Inheritance of recombining and non-recombining portions of the genome for three generations

The Y chromosome (small chromosome symbol) and the mtDNA (circle) of the son are inherited from one grandparent, the father's father and the mother's mother, respectively. The autosomes (large chromosome symbols) are inherited from all four grandparents and mixed because of recombination. The different characteristics (hereafter briefly described) and the different patterns of inheritance of these three parts of the DNA make them useful in several applications.

1.2 Mitochondrial DNA

The mtDNA (figure 1.2), is a circular double stranded DNA molecule of about 16.5 kb in lenght, whose entire sequence is known (Anderson et al. 1981a; Andrews et al. 1999). It presents a heavy strand (relatively rich in G bases) and a light one (relatively riche in C bases). The control region contains two hypervariable regions, usually assaied for variability. The mtDNA is not contained in the nucleus, but within the mithocondrial-cytoplasmatic organelles in which the energy-genereting process of oxidative phosphorylation takes place. The mitochondria originated as endosymbiontic bacteria in proto-eukaryotic cells about 1.5 billion years ago, and provided energy generation. Many features of the mitochondria are similar to those of the bacteria: the circular than linear genome, the absence of histones, the discrete origins of replications (e.g. the absence of introns genes, the absence of dispersed repeats and the very little amount of inter-genic DNA). The most striking distinction between mitochondrial and nuclear DNA is the different genetic code: in mammals, five codons have different specificities in mtDNA compared to nuclear genome.

The 37 genes that are present in the mtDNA of modern humans enter in either oxydative phosphorylation pathway or mitochondrial protein synthesis. During the evolution, the other genes related to mitochodrial function have been transferred to the nuclear genome.

The number of mitochondria in a cell depends on its type and energy-request. Cells requiring a lot of energy, as nerve and muscle cells, contain thousand of mitochondria, each containig 2-10 copies of mtDNA, while other cell types may contain only a few hundred. Since the oocytes contain around 100,000 mitochondria, each containing a single mtDNA molecule, and sperm contain only about 50-75, the paternal contribution to the zygote's pool of mtDNA is expected to be relatively small. However, there is evidence that its contribution is actually null, and the mtDNA from different lineages could, in principle, happen, since the necessary complexes are present. This element together with some features of the pattern of polymorphic sites within the mtDNA molecule led to much debate over whether mtDNA does indeed recombine.

The characteristic of the mtDNA is that it is not subject to recombination and evolves for sequencial accumulation of single nucleotide substitutions. Consequently all its variability is the product of this accumulation along the maternal line (Birky 1983; Giles et

al. 1980). Moreover, the mtDNA has a high rate mutation thanks to the lack of reparation enzymes: this determines the fast accumulation of nucleotide substitutions making possible to distinguish populations that separated in relatively recent times. Mitochondrial sequences with a common origin can be grouped under the same haplogrup. Different populations sharing one or more mitochondrial haplogroups can indicate that their divergence is relatively recent.

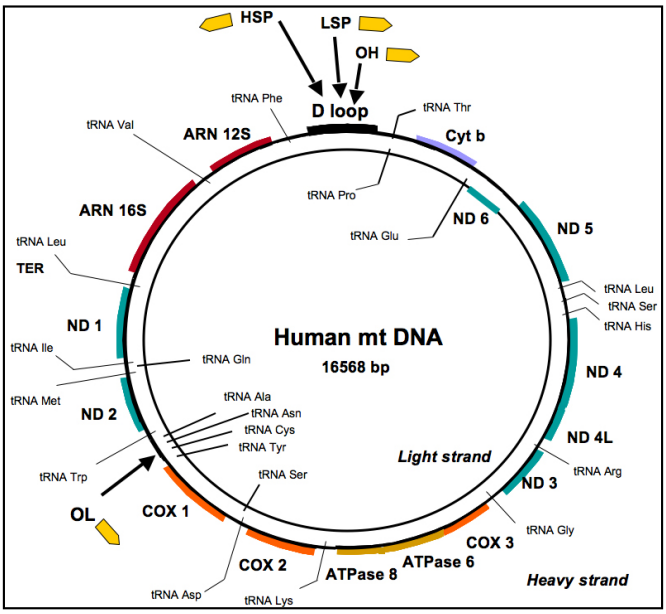


Figure 1.2 Human mitochondrial DNA

1.2.1 Inheritance

The mtDNA is inherited in a non-mendelian way, and the fact that it (and the mitochondria) is maternally inherited and is not recombing, has been considered for a long time a dogma, even though in others organisms as plants, fungi, mussels or fish we do have mtDNA recombination. Giles et al (1980) analyzing sequence polymorphisms by RFLPs, observed that the descendants only had the maternal variant. Later studies confirmed that sperm mitocondria is selectively destroyed in the oocyte and that the paternal mtDNA is marked by ubiquitinacion during spermatogenesis for its later destruction in the oocyte.

Recent studies questioned the dogma: Schwartz et al (2003) observed in a case of severe intolerance to exercise, that patient mtDNA was predominantly of paternal origin. Zsurka et al (2007; 2005) carried out similar studies in families in which several members presented heteroplasmy in skeletal muscle; Kraytsberg et al (2004) found recombination in a patient between the maternal and the paternal genome, of approximately 0,7% of the total mtDNA of its muscular tissue. Nevertheless, all these studies do not explain why recombination signals do not appear in the mtDNA philogeny or why the majority of heteroplasmies occurr generally in only one or two profile positions, instead than in more positions, which would be incompatible with a model of patrilineal mitochondrial contribution (Bandelt et al. 2005; Bandelt et al. 2004d). The reanalysis of many studies in which recombination evidences had been observed, allowed to verify errors in the data as well as in the statistical methods used (Pakendorf and Stoneking 2005), bringing to some authors' rectifications (Hagelberg et al. 1999), and, in some other works, mtDNA mosaics could be due to contamination or mixes of different samples and not to recombination (Bandelt et al. 2005).

Therefore, it seems very probable that the unique case of mtDNA paternal inheritance detected in humans is a failure of the normal recognition and elimination process of the paternal mitochondrial genome molecules, which is an extremely rare phenomenon. Thus, at least at the moment, the mtDNA maternal inheritance can still be considered the rule, so that all the individuals of a same maternal lineage exhibit the same mtDNA sequence (unless in exceptional cases where heteroplasmy segregation exists in some individual of the lineage, or punctual mutations at a germinal level). This way of uniparental inheritance and its haploid nature is one of the great mtDNA advantages, since this allows researchers to follow the lineages backwards in time, looking for the maternal ancestry of a population, avoiding therefore the effects of the biparental inheritance and recombination (Pakendorf and Stoneking 2005).

1.2.2 Mutation rate

Mutation in mtDNA has two unsual aspects, i.e. its high average mutation rate with respect to nuclear DNA, and the manner in which mutations pass from one generation to the next.

Mutations occuring in mtDNA are largely base substitutions and deletions that can involve much of the mtDNA molecule; delections are the causes of maternally inherited mitochondrial diseases, while many base substitutions appear to be neutral. Evolutionary genetic studies concentrate on these, but also consider other apparently neutral variations, such as the lenght change in a poly(C) tract in the control region, and the '9 bp deletion', a change in copy number of 9 bp tandem repeat motif from two copies to one copy.

Early comparisons of human and other primate DNAs indicated that the base substitutional mutation rate in mtDNA is about 10 times higher than the average rate in nuclear DNA (Brown et al. 1979), leading to a very large number of different sequences in human population, which is possibly due to the reduced fidelity of mtDNA replication and/or repair systems. In the mitocondria, more than the 90% of the oxygen entering the cell to provide the energy it needs, is consumed, and this metabolic activity is generating a great amount of free radicals with mutagenic effect.

The observed mutation rate throughout the 16.5 kb molecule is nonuniform, with relatively low rate in the coding regions, and a relatively high rate in the control region, which does not encode proteins or RNAs. At some positions within the hypervariable segments (HVS-I and HVS-II, sometimes also known as hypervariable regions, or HVRs) of the control region, mutation rate is so high that mutations can often be observed in pedigrees, and comparisons between the pedigree rate and the rate calculated in other ways, have caused controversy.

A good estimate of mtDNA mutation rate is important for human evolutionary genetic studies. Great interest has focused on the mutation rate of the control region, since this is the part of mtDNA most widely used in population studies. MtDNA mutation rate are often expressed as base sostitutions 'per site per million year', rather than the 'per base per generation'.

Mutation rate depends on how fast mutations appear and fix in the different lineages, and allows to introduce a temporal scale in molecular evolution and, thus, to estimate the TMRCA between two lineages, that is the time of divergence among them.

The evolution rate is based on two processes: frequency with which new mutations arise in the molecule, and probability these new mutations fix in the population (fixation rate). The way a mutation occurs and fix in the maternal line is important not only in mitochondrial disease investigations, but also in the human population studies (mtDNA

sequence diversity indices are used to date demographic events), and for identification and criminal cases in forensic genetics. The Coding Region mutation rate is 2-4% per million years (Cann et al. 1987; Stoneking et al. 1992), except the control region, whose rate is higher. It is difficult to have a unique mutation rate value, since it is a very controverted subject and a consensus does not exist (Pakendorf and Stoneking 2005). In order to estimate the mtDNA mutation rate, different approaches can be used: through pedigree or philogenetic studies, depending on the system used, but different authors obtained different values. Investigations based on familiar studies, gave values quite different between them: Howell et al, (1996) obtained a mutation rate of $1.08x10^{-5}$ per site per generation, whereas Bendall et al (1996) put this value between $1.2x10^{-6}$ and $2.7x10^{-5}$ and Parsons et al (1997) in $2.7x10^{-5}$. Phylogenetic comparisons, based on interspecific or intraspecific comparisons, provide more similar values between each other, and inferiors to the ones obtained by direct mtDNA familiar mutations observation (Stoneking et al. 1992; Ward et al. 1991). Now the debate concentrates on how this can reflect real events and which value should be used for population history studies (Pakendorf and Stoneking 2005). In order to understand these differences between estimates based on pedigrees and on philogenetic studies, it is important to consider that the level of molecular stability varies for the different sites throughout the mtDNA molecule, with positions within the control region known as "hot spots" or "hot spot mutations", whose mutation rate is 4 or 5 times higher than the average (Pakendorf and Stoneking 2005). Because the generational values include very small periods, with these kind of positions the mutation rate is overestimated, on the contrary the phylogenetic studies include wider evolutionary periods and the values obtained would reflect average values of all the mtDNA variable positions throughout a lineage, and this esteem would be correct if considering the hot spot recurrent mutation rate (Sigurğardóttir et al. 2000). Although it is important to consider the mtDNA mutation rate, practically the site-specific mutation informations are more important.

Some of the most commonly used mutation rate in the last years is $1.79x10^{-7}$ base sostitution per nucleotide per year for 276 bp of the HVS-I obtained by Foster and collegues (1996), using a network method calibrated by the assumption that the expansion of haplogroup A2 happened 11,300 years ago. Another one was presented by Ingman et al (2000) and it is based on coding region, assuming a human-chimp species split at five million yeras ago, yelding an estimate of $1.7x10^{-8}$ base sostitution per nucleotide per year. Other two

values are $1.26x10^{-8}$ (Mishmar et al. 2003) and $3.5x10^{-8}$ base sostitution per nucleotide per year (Kivisild et al. 2006) calibrated assuming that the split between humans and chimpanzees happened 6.5 Mya.

Soares et al (2009) obtained then a new value for the substitution rate for the entire molecule of $1.665x10^{-8}$ base sostitution per nucleotide per year, considering the divergence human-chimpanzee at 7 Mya. They calibrated the mtDNA clock considering this divergence but adding a correction for selection and saturation effects.

1.2.3 mtDNA applications in forensic genetics

1.2.3.1 *General aspects*

Until now, mtDNA analysis was the only solution for degraded samples when nuclear markers were failing. Due to its high number of copies per cell and to its structure, it is much more resistant to degradation than nuclear DNA, so that even 200 bp fragments can be successfully amplified in samples that completely fail for STRs, and even with ancient DNA samples. In addition to these first characteristics, also its high mutation rate leads to the genesis and accumulation of a great number of variants in the human gene pool. This diversity provides a great support for paternity tests and biological remanis identification. Moreover, mtDNA molecule is transmitted as a unique locus continously through maternal lineages. This fact makes possible the biological remain identification in absence of direct relatives, through the comparisons with individuals belonging to the same maternal lineage.

Finally, being SNPs, nowadays, the polymorphic markers mainly studied, mtDNA contribute to forensic genetics is dued to the facility for multiplex typing and automatization, the simplicity, the robustness and, generally, the advantages that accompany SNPs laboratory methodologies.

The use of mtDNA in the resolution of judicial forensic cases is relatively recent, but nowadays, it is considered a routine analysis in many world laboratories.

Databases. A great number of publications in forensic literature alert of the existance of high incidence of errors in mtDNA databases. Nevertheless, this is not a topic affecting only the forensic field (Bandelt et al. 2004a; Bandelt et al. 2004b; Yao et al. 2004), but also other

areas of investigation, like population genetics and clinical studies (Bandelt et al. 2001; Kong et al. 2006; Salas et al. 2005).

When in a mtDNA database study a high number of transversions are observed, this should be a reason of suspect on the dataset, since generally transitions are much more frequent than transversions in all the genome. In the same way, insertions and deletions (excluding the positions in the homopolimeric part of the region control) are not so frequent. A good system to detect the majority of these mistakes is to use a phylogenetic approach (Salas et al. 2007), although with this method, only a limited number of errors depending on the haplotipic composition of the studied population can be detected, since some phylogenies are better defined than others.

The impact of mistakes on mtDNA databases, is different depending the studied fields. In population genetic studies, errors will affect questions related to the mtDNA evolution. The debate on mtDNA recombination is known, but also the great impact that published errors had at the time and, in some cases, the later rectifications (Hagelberg et al. 1999). They can also affect phylogenetic and demographic analyses, altering the esteem of some populational events. Nevertheless, the consequences in the forensic field can be much more serious, like false exclusions and statistical evaluation distortions (Salas et al. 2007). In the forensic context, databases are absolutely necessary for mtDNA evidence interpretation. The importance of the evidence depends on the frequency a certain profile appears in the reference population, but often the number of samples in a database is small in relation with the great variability of the population; sometimes the sample is not representative of the population, with no suitable haplotipic frequencies. In order to consider these frequencies in a proper way, great databases are needed. The greater mtDNA public database is the one of the Scientific Working Group on DNA Analysis Methods (SWGDAM), containing around 5,000 profiles. The problem is that these profiles are divided in small data groups, limiting the haplotypic estimate use. The new mtDNA database of EMPOP (http://www.empop.org) is organised in order to avoid, as far as possible, the addition of erroneous profiles, trying to consider all the world populations, even if, at the moment, the majority of the profiles included belong to west Eurasian populations. Another important problem to consider is related to the use of the mitocondrial databases in a criminal case, assuming that the database is representative of the regional population, and normally corresponds to a geographic location or to an ethnic group.

1.2.3.2 *mtDNA disadvantages in forensic analysis*

On the other hand, mtDNA has two important disadvantages dued to its exclusively matrilineal inheritance: on one hand, the limitation to use mtDNA in cases where the involved individuals are exclusively maternally related and, on the other, the fact it does not recombine and it is transmitted like a unique haplotipic block. The mtDNA non-mendelian way of inheritance implies taking certain precautions, necessary for the use of mitochondrial genome polymorphisms in forensic genetics. The difficulty is based on the fact that the used markers do not present independent inheritance. According to this, an identification profile calculation probability cannot be done by individual allele frequencies for each variant, but by haplotype frequency in the given population. All the mtDNA polymorphisms, could only be considered as a unique locus, limiting in this way their informativness, assigning a profile not to an individual but to an extensive lineage.

1.2.4 mtDNA variability in human populations studies

The different human populations present a different mtDNA variants distribution. This is the result of the neutral evolution processes on human mtDNA, after the initial exodus from Africa started 100,000 years ago (Cann et al. 1987).

The unique characteristics of the human mtDNA make this molecule useful in historical and human populations evolution studies (Pakendorf and Stoneking 2005). Phylogeography studies the geographic distribution of the different mitochondrial lineages. Analysing the present lineages, history can be reconstructed following the coalescent principle, and so, going back to the unique, common maternal ancestor. The evolutionary scale can be analysed at different levels: within a same population, between populations of a the same species or among different species. In order to infer phylogenetic relations, the homology criterion has to be applied as a basic norm, so that the compared character has to be homologous, that is to say, it must be shared by ancestry or be identical in the line. The phylogenetic reconstruction can be done with different procedures, but the most used is the "maximum parsimony" or "minimum evolution" one: evolution happens parsimoniously so that the nucleotide changes for all the positions have the same probability and the reconstruction assumes the simplest explanation for the information, which means that the

chosen tree is the one that requires the minimum mutational number events to explain the distribution of the characters between the studied individuals. Most of the studies on mtDNA variation in different human populations have been realised by two methods: direct sequencing of the control region or amplification of the complete mtDNA molecule in nine 1,500 to 3,000 bp overlapping fragments or with a set of 14 restriction enzymes (Chen et al. 1995; Torroni et al. 1992). The obtained haplotypes can be classified by parsimony analyses in different groups, sharing the same point mutations. The different groups are called haplogroups and are specific for certain population groups, giving a great information on the interpopulationals connections. Despite the control region high average mutation rate, some positions are decisive for the haplogroup determination.

In the last few years, with the high throughput technologies development, the complete genomes sequencing has become common even in population studies (Herrnstadt et al. 2002; Ingman and Gyllensten 2001; Yao et al. 2006). The phylogeny based on the complete genome data, offers a better resolution than the one obtained only by sequencing the complete control region. A good strategy combines the control region data and the coding region positions, traditionally analyzed with RFLPs, and in the last few years with other technologies, like the minisequencing SNaPshot technique (ABI PRISM® SNaPshot™ Multiplex System, AB) (Álvarez-Iglesias et al. 2007; Brandstätter et al. 2006; Quintáns et al. 2004).

The results interpretation of mtDNA analysis had a great impact and consequences in the scientific world, stating the African origin of the mtDNA molecule, and consequently, of human populations. Cann et al (1987) considering the results obtained for the genetic studies on populations from Asia, Australia, New Guinea, Europe and Africa, formulates the well-known "Mitochondrial Eve" hypothesis. The three main aspects of this hypothesis are: (I) all the current mtDNA types go back to a unique ancestor, (II) this ancestor lived in Africa, and (III) probably it lived about 200,000 years ago. This interpretation has been tremendously controversial essentially because of the use of certain phylogenetic reconstruction methods. Despite that, at the moment the African hypothesis of the origin of the modern man has been supported by a great number of mtDNA but also nuclear markers investigations. This hypothesis, later demonstrated by other authors (Chen et al. 1995; Horai and Hayasaka 1990; Ingman et al. 2000), is based on the coalescent principle, according to which, assuming that there was a unique origin for all the living organisms, all the mtDNA

(or nuclear) present variations are proceeding from a unique ancestor, existed in previous generations. In the mtDNA case, being of maternal unilinear inheritance, the lineages reconstruction is much more simple. This ancestor was feminine (Eve), it was not the unique living individual, but a population member, and it only represents the point from which all the lineages gather.

MtDNA not only contributed to the explication of the humans origin, but also to the human migrations reconstruction: the New World and the Pacific colonization, the initial migrations to New Guinea and Australia and the settlement in Europe (Pakendorf and Stoneking 2005). Throughout history and during all these population movements, mutations started accumulating sequentially in the mtDNA lineages starting from founder sequences. These maternal lineages diverged as the human populations colonized the different geographic regions all over the world, reason why, many mitochondrial haplogroups are continent-specific (Figure 1.3). Haplogroup L, with L1, L2 and L3 sub-groups, is typically African, whereas M and N haplogroups originated in east Africa from L3, dispersed towards Eurasia and the New World. Haplogroups H, I, J, N1b, T, U, V and W are characteristic of Caucasoid populations, haplogroups A, B, C and D are specific for Asia and the New World, and haplogroups G, Y and Z predominates in Siberia.

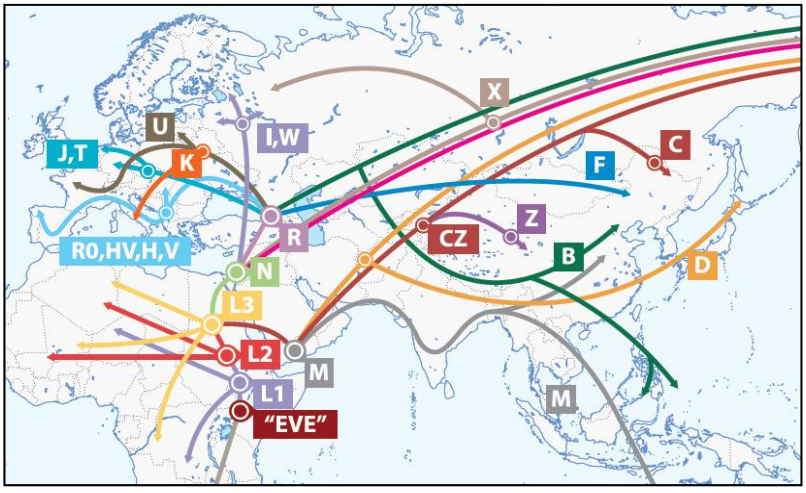


Figure 1.3 mtDNA Human Migration

Nevertheless, mtDNA represents only one locus and partially reflects the matrilineal demographic populations. The history written using only one locus cannot faithfully reflect a population history, reason why the studies realised using mtDNA variability must be complemented with other markers data: Y chromosome, and ideally, also with autosomal data (Pakendorf and Stoneking 2005)

RFLPs studies (Chen et al. 1995; Torroni et al. 1992) revealed that the majority of the African, Asian, European and American mtDNAs are defined, at least, by continent specific polymorphisms. Their high frequency within the main continental groups allow to infer the ethnic and geographic origins of the modern groups, suggesting their later derivation from ancestral populations. Complete mithocondrial genome studies (Finnilä et al. 2001; Maca-Meyer et al. 2001) also support the "Out of Africa hypothesis", since the African population presents a higher sequence diversity, as a result of the great and constant population size and a longer genetic history. Non-African populations, in fact, present a "star-like" philogeny, suggesting a bottle neck event and a subsequent but only recent population expansion.

The TMRCA for modern humans varies considerably between the different studies, since the substitution rates used are different and were progressively improved: 171,500±50,000 (Mishmar et al. 2003), 198,000±19,000 (Ingman et al. 2000; Mishmar et al. 2003), 204,700±22,100 (Macaulay et al. 2005), 192,400±41,000 (Soares et al. 2009).

Dispersion routes and colonization of the world. The analysis of the mitocondrial genome variation has also contributed to the knowledge of the human populations dispersion routes. The weight of the three main hypotheses changed over the years, because of the new genetic data results and archaeological findings.

- North route: this is the dispersion route through North Africa and Levant, approximately 45,000 years ago, and from here towards Europe and Asia. For that reason, all the present mitocondrial haplotypes should coalesce to the original mtDNA founders arrived to these regions.
- South route: this hypothesis proposes a dispersion by the south coast of Asia, arising from the east of Africa, approximately 60,000 years ago.

• Two dispersion waves: according to this hypothesis, colonization should be explained by the two previous routes, with a first (about 60,000 years ago) wave of dispersion through the South coast of Asia, giving rise to the mtDNA haplotypes of the south and east Asia, and the Australian Continent, and later (45,000 years ago) another trend of dispersion by North Africa to the Middle East, from where migratory movements took place towards the north of Asia and Europe (Maca-Meyer et al. 2001). The mtDNA also has contributed to the knowledge of the colonization of the New World (Perego et al. 2009; Torroni et al. 1992), the initial migrations to New Guinea and Australia (Ingman and Gyllensten 2003) and the settlement in Europe (Torroni et al. 2006).

1.3 Y Chromosome

The Y chromosome has the sex determination role, it is haploid, and it is transmitted from father to son. This acrocentric chromosome represents only 2% of the human genome, and has an approximate lenght of 60 Mb, shorter than its homologous. It escapes recombination for most of its lenght (> 90%), known as the NRPY (or NRY) or as the MSY region. In fact the Y chromosome actually recombines with the X chromosome in male meiosis only in specialized regions, where sequence identification with the X is preserved (Figure 1.4). Sequences within these regions can be inherited from both, father and mother; thus, they are referred to as pseudo-autosomal. The PAR2, at the tips of the long arms of the X and Y chromosome, is scarcely important in the chromosomal segregation and its acquisition is evolutionary recent in humans. On the contrary, PAR1, which is a 2.6 Mb region at the tip of the short arms, reflects an ancient origin of the mammalian sex chromosome as a pair of homologous autosomes, and it is the site of an obligate event in male meiosis.

The Y chromosome consists of two regions: an heterocromatic one, mostly localized on the long arm, with DNA ripetitive satellites at the distal part (DYZ1 and DYZ2), and an euchromatic one with less condensed DNA and proteins. The 23 milllion bp of the euchromatic part represent the MSY region and are composed of three classes of sequences:

1. X-transposed, which are 99% identical to human Xp21 (X chromosomal) sequence. These regions do not participate in XY crossing-over. Their combined lenght is 3.4 Mb, and only 2 genes having omologues on Xq21 were identified.
2. X-degenerate, remnants of the ancient autosomes from which the modern human X and Y chromosomes evolved. They contain single copies of genes or pseudo-genes homologues to 27 different X-linked genes. The sequence identification with their homologues is between 60% and 96%.
3. 'Ampliconic regions', including large regions where pairs of sequences have an identification higher than 99.9%. They map on 7 sequence segments dispersed over the long arm, in the eucharіotic region and the proximal part of the short arm. The combined lenght reaches 10.2 Mb. These sequences show the highest gene density (both coding and non-coding genes) among the 3 regions. An

alphoid region made block of satellite DNA is present in the centromeric region (Quintana-Murci et al. 2001)

Most of the genes on the Y chromosome are located in proximity of the transition between the pseudo-autosomal region of the short arm and that of the centromere, and in proximity of the transition from the latter to the heterochromatic region of the long arm (Yq).

The Y chromosome cannot contain important genes for the survival that are not shared with the X chromosome, since it would be absent in women. Consequently, the Y chromosome is extremely gene-poor: the non-pseudo-autosomal euchromatin, 23 Mb in lenght, contains only 27 genes (1.2/Mb; (Skaletsky et al. 2003)), while the X chromosome presents 717 genes in 160 Mb (4.5/Mb).

The Y chromosome has peculiar evolutionary characteristics with respect to the other nuclear loci: it has higher average mutation rate, higher between-species sequence divergence, but lower within-species diversity. The higher average mutation rate is due to the passage through the male germ-line, more mutagenic than the female germ-line. This is why the sequence divergences between hominoid species are larger for the Y chromosome than for other loci. The smaller effective population size (the number of individual in a population who contributes to the next generation) makes it subject to drift and so to a smaller diversity within a species. The increased genetic drift also leads to large differences between populations, making the Y chromosome the most geographically informative locus in the genome (Jobling et al. 2004), as from present knowledge (Wilder et al. 2004).

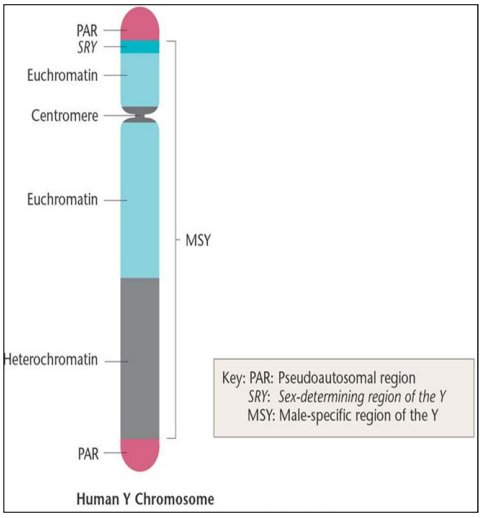


Figure 1.4 - Y chromosome

1.3.1 Inheritance

One of the most interesting Y chromosome characteristics is its way of inheritance: it only passes from father to son, and so, it is male specific. The recombination absence practically in all the chromosome allows to pass from one generation to the other like an identical haplotipic block. Mutation is the unique way producing haplotipic diversification, being this process exclusively intra-allelic. In addition, using low mutation rate polymorphisms, like SNPs, a philogeny could be easily reconstructed. It is necessary to clarify, due to its inheritance form, that the effective Y chromosome population size is four times lower than any autosomal chromosome and three times smaller than X chromosome. The reason is that each couple descendant can inherit 4 possible autosomal chromosomes (two from the father and two from the mother) and 3 possible X chromosomes (two from the mother and one from the father), whereas for the Y chromosome there is only one possibility (the father Y chromosome), in case of having a male progeny (Jobling and Tyler-Smith 1995). This Y chromosome reduced size, makes it more susceptible to undergo the effects of the genetic drift, bringing random changes in the haplotipic frequencies from one generation to the other, speeding up in that way the chromosome groups differentiation between different populations. The effects of the genetic drift turn Y chromosome in one of the most

interesting tool to investigate past population events, although sometimes, drift can produce very fast changes in the haplotipic frequencies making difficult the interpretation of certain quantifiable aspects of these passed events, like for example, admixing proportions between populations or past demographic changes (Jobling and Tyler-Smith 2003). In addition, Y chromosome inheritance patterns can influence genetic diversity depending on the migratory behaviours. The different sex migration behaviors are obvious: males migrated more than females in the modern world; examples of intercontinental migrations, besides recent historical documents, show that explorers, traders and soldiers have been almost exclusively men. Nevertheless, considering the impact of migration in genetic diversity, we must examine not only the migration patterns on a great scale, but also, the local migrations. Talking about local migrations, we have to consider two patterns: 1) Patrilocality, when the family unit locates close or in the same male family area; 2) Matrilocality, the inverse phenomenon, when the male move to the female family locality.

Patrilocality occurs approximately in the 70% of the world-wide population (Burton et al. 1996). This implies that mtDNA (exclusively of maternal inheritance) is mainly moving around between populations in each generation, whereas the Y chromosome remains within each population. It is necessary to consider that migration entails a reduction of the population differentiation and according to some works (Kayser et al. 2001a; Seielstad et al. 1998) Y chromosome presents a higher degree of genetic differentiation between populations than mtDNA, thus suggesting a greater feminine migration rate in the past. These works demonstrate that these small migrations are more important, from the genetic and geographic diversity patterns point of view, than the migratory processes impling great distances. These migratory characteristics together with the reduced Y chromosome population size can help interpreting the low variability found for this chromosome.

1.3.2 Mutation rate

Y chromosome polymorphisms. Eukaryotic cells present a lot of DNA repeated sequences. These DNA repeated sequences have all type of sizes and differentiate by the repeated length units and by the number of times that a unit is repeated between different individuals. So, it is possible to find: 1) great size repetition units (from several hundreds to several thousands bases), like satellite DNA, close to the chromosomal centromeres; 2) smaller repetitions (10-

100 bases), denominated minisatellites DNA or VNTR; 3) and finally, the smallest size repetitive units (2-6 bases of length) called DNA microsatellites, SSRs or STRs. Besides this type of tandem repetitive DNA, another type of dispersed repetitive DNA is distributed along all the genome (Smith et al. 1987):

- SINEs: the size of the repetition unit is between 100 and 500 bases. Alu are the most abundant elements in the human genome (around a million copies). The SINEs elements constitute 13% of the human genome and probably these elements are generated by intrachromosomic genic conversion.
- LINEs: the size of the repetitive unit is usually greater than 500 bases. In the genome there are about 900,000 LINEs making up the 23% of the human genome: the majority belongs to LINE-1 family (L1; only 50 elements are functional genes). Even in this case, these elements are generated by intrachromosomic genic conversion.

Another type of variability is generated by point mutations, impling a nucleotide change (SNPs), or the insertion-deletion of one or several nucleotides (Indels). This variability becomes a polymorphisms when one of the 2 possible alleles (sometimes can have more than two allelic possibilities) have a frequency of 1% or more in the population.

Binary polymorphisms (SNPs). It is a simple polymorphism based on the change of a nucleotide base for another. These markers are some times referred like UEPs, because of their very low mutation rate, that can be approximately 10^{-8} mutations/generation (10^{-3} approx in STRs) (Kayser et al. 2000; Nachman and Crowell 2000; Thomson et al. 2000). A part from the SNPs, Indels are another important type of binary markers, and consist of insertions/deletions of one or several nucleotides. Basically, the processes generating this type of polymorphism are: errors in the nucleotide incorporation during the replication, and mutagenesis caused by the chemical modification or by physical bases damages. The frequency is 1 each 100-500 bases in the human genome and the majority occurs in the non coding regions, having no known effects on the individual phenotype. These SNPs are very useful in population genetics and evolutionary studies, but also as identifing markers in genetic diseases, like linkage family studies, observing the linkage disequilibrium in isolated populations, on heterozigosity loss studies in tumors and in association analysis between

patients and controls. Even if SNPs, singularly, are less informative than others more frequently used genetic markers, they are abundant and spread in all the genome, with a great automatization potential.

Microsatellites or STRs. The STR loci are highly polymorphic, and belong to the length polymorphisms group. The microsatellites are loci whose sequence is formed by tandem repetitions of a same 'nucleus sequence' (between 7 and 2 nucleotides), being the variation the number of this nucleus repetitions, with a total extension that varies typically between 100 and 400 bp. These markers, in addition, have an ample distribution throughout the genome, with an average value of an STR every 5,000 or 10,000 bases distributed along the noncoding DNA, as well as in genic and intergenic regions. Due to the high variability of these regions, a typical STR will present an average of more than ten variants in the population, depending also on the locus own structure and the tandem repetition of short DNA sequences (Bässler et al. 1999).The number of repetitions of an STR can vary among individuals (the most interesting are the ones having between 10 and 30 repetitions), reason why they turn into a very interesting tool in human identification.

Actually, ther are thousands of identified polymorphic microsatellites in human DNA. They are scatterd in all the chromosomes. Microsatellites have an important role in forensic genetics, particularly in sexual aggression or paternity tests, mainly when the presumed biological father is not available (Jobling et al. 1997; Kayser et al. 1997a); they are also very useful in genealogical investigations (Jobling 2001a) and in evolutionary studies, as male lineage markers (Jobling and Tyler-Smith 2003; Kayser et al. 2001b; Stumpf and Goldstein 2001). Basically the Y-STRs are used with 2 purposes (Kayser et al. 2004): first to distinguish lineages, in this case the number of markers and their variability will determine the discrimination level, and then to provide correlation informations between lineages.

In evolutionary studies up to 16 Y-STRs are used, and in some populations the majority of the individuals share the same 16 Y-STRs haplotype (Hedman et al. 2004; Zerjal et al. 2003). In exceptional cases, also father and son can differ in their Y-STRs haplotype because of a mutational event (Kayser et al. 2000). That possibility must be considered in statistical calculations in forensic and paternity cases (Kayser and Sajantila 2001; Rolf et al. 2001). In order to have reliable results, it is important to use a great number of microsatellites and to know their haplotypic frequencies, as well as the mutational characteristics of each locus (Stumpf and Goldstein 2001).

The similar allele sizes for each locus inspired the "stepwise mutation" mechanism hypothesis or model (Ohta and Kimura 1973), where the replicative slippage could also be implied (Levinson and Gutman 1987; Stephan 1989; Tautz and Renz 1984; Weber 1990). Microsatellites studies on genetic diseases (Mahtani and Willard 1993; Petrukhin et al. 1993; Weber and Wong 1993) show that more than 90% of the mutations are dued to the increase/diminution of a single repetitive unit with respect to the initial allele (Calafell et al. 1998; Cooper et al. 1996), especially for the 2 to 4 bp repetitive units (Jin et al. 1996). This model probably does not consider all the variation types found (Deka et al. 1995; Shriver et al. 1993) and so a new different model with two mutational phases was proposed (Di Rienzo et al. 1994): on one hand we would have the mutations with a unique repetitive unit variation, the most frequent, and on the other hand, the mutational changes in which several repetitive units are implied, considering that length variations on greater scale are less frequent.

Probably, the chain slippage during replication is the mechanism generating microsatellite mutations (Brinkmann et al. 1998; Heyer et al. 1997; Kayser et al. 1997b; Levinson and Gutman 1987; Schlotterer and Tautz 1992; Strand et al. 1993). Slippage is an intrachromatic phenomenon that takes place during the replication process, when a DNA polymerase displacement takes place on the template strand of the DNA repeated region and the enzyme reassemblies in an ahead position. As a result of this scrolling, DNA polymerase incorporates the complementary bases, causing an erroneous pairing between both DNA filaments. Depending on which filament, parental or new synthesis one, the error is produced, the result is respectively the delection or insertion of new repetitive units (Sia et al. 1997).

A good esteem of the Y-STRs mutation rate is a fundamental requirement in order to date correctly the Y chromosome lineages defined by SNPs (Bianchi et al. 1998; Zerjal et al. 1997), as well as for the data interpretation in paternity tests and in forensic cases (Jobling et al. 1997; Kayser et al. 1997b). At the moment, the number of Y-STRs mutation rate studies is scant and are always focused on the same STRs. One of the widest study is from Gusmão et al (2005) where 16 Y-STRs are used and the mutation rate value is 1.998×10^{-3}, that is more or less the same values also obtained by Kayser et al (2000) and Dupuy et al (2004), 2.8×10^{-3} and 2.31×10^{-3}, respectively. Heyer study had practically the same results (Heyer et

al. 1997), using 7 Y-STRs, and obtaining a value of $2x10^{-3}$ (Weber and Wong 1993). On the other hand, Foster (2000) first realizes the mutation rate evolutionary importance for populational events dating. Using 6 Y-STRs and 5 SNPs, based on phylogenetics networks, the mutation rate esteem obtained was of $2.6x10^{-4}$ mutations/20 years, almost 7 times lower than Heyer et al (1997). Similar results were obtained by Zhivotousky et al (2004), with an average mutation rate per repetition per STR of $6.9x10^{-4}$/25 years, using both STRs and SNPs. All these data complicate the work at the time to choose an appropriate mutation rate for investigations, since as Zhivotousky (2005) proposes, a discrepancy is possible between the mutation rate existing from the genealogical and evolutionary points of view.

1.3.3 Y chromosome applications in forensic genetics

The Y-SNPs are mainly used in molecular anthropology for evolutionary studies. Nevertheless, although they present an inferior discriminatory power respect STRs, they offer great advantages especially in the forensic routine. The possibility of being analyzed in small amplicons, and so the possibility to work on degraded samples, the automatization improvement, the high through-put platform analysis, and finally the Y-SNPs haplogroups geographic distribution, makes them really helpfull. The last characteristic is a tool to predict the possible paternal lineage geographic origin, from biological remains found in a forensic crime scene. In a criminal survey, the statistic interpretation in case of exclusion, is direct, but in case of matching profiles, the marked geographic distribution makes the necessary statistical corrections (Jobling 2001b). Although Y-STRs are the election markers in these cases, Y-SNPs are the markers used in complicated cases with masculine and feminine admixture, for the utility of the complementary informations provided (Sánchez et al. 2004).

Nowadays, the co-amplification of different Y-SNPS by a multiplex technique allow the characterization of the main phylogenetic tree clades and a geographic discrimination, just in a simple reaction (Brión et al. 2005a; Brión et al. 2005b; Sánchez et al. 2003). New typing methods and techniques arise continuously (moreover the user will prefere a technique or another depending on the case), but generally in forensic genetics the great capacity and high precision multiplexes use, is preferred (Sobrino et al. 2005).

1.3.3.1 *Y chromosome in forensic applications*

Within all the STRs, the Y chromosome specific are an important tool for the majority of the forensic genetic laboratories, although they do not allow the same level of (Brión et al. 2005b)resolution in identification like the autosomal STRs, because a Y-STRs haplotype is a non-recombining lineage that can be shared by many individuals (Gusmão et al. 2006). The Y-STRs are usually used in paternity tests when we have a son and not the father, but instead, for example, some male individuals from the paternal line: a brother of the father, the paternal grandfathers of the son, etc (Gill et al. 2001; Kayser et al. 1997a; Roewer et al. 2000). In addition they are especially usefull in sexual aggression cases since the Y-STRs will allow the masculine component identification from corporal fluids, when a male and female cells admixture exists, as well as the determination of the number of males implied in multiple sexual aggressions.

The constant Y-STRs discovery is beginning to suppose an unambigous nomenclature, because not all the laboratories equally named the same STR and their allelic variants (Ayub et al. 2000; González-Neira et al. 2001; Gusmão et al. 2002; White et al. 1999). So, the ISFG DNA commission created some norms to establish a common nomenclature, in order to allow the communication and the data exchange between the different forensic laboratories (Gill et al. 2001; Gómez and Carracedo 2000; Gusmão et al. 2006). The ISFG DNA commission publishes regularly new recommendations related to the Y-STRs application in the human identification, besides many other recommendations for other polymorphisms and procedures in the forensic genetics field (Gill et al. 2006; Prinz et al. 2007). When a match is found between the Y-STR haplotype, for example, from a blood spot on the victim and the suspected, we need to know the haplotype frequency in the population in order to calculate the matching probability. This implies the necessity to create large databases with complete Y-STRs haplotypes of the population, in order to calculate reliable haplotipic frequencies (Hammer et al. 2005; Lessig et al. 2003; Roewer et al. 2001). A forensic databases should have three main characteristics (Roewer et al. 2001):

- polymorphisms able to discriminate between the majority of no related lineages in a given population.
- representative ethnic-geographic structure data of the considered population.

-.a data size that allows to consider with precision the frequency of the rare haplotypes.

Nowadays one of the wider database is the YHRD, initially containing only 8 Y-STR, constituting the "minimum haplotype" used in forensic genetics and including: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a/b (Pascali et al. 1999). This database is the result of an international project born in 1994, that was looking for the Y-STRs optimization possibilities in the forensic field. It is available for free online (http://www.yhrd.org). Recently 8 new Y-STRs were added: DYS438, DYS439, DYS437, DYS445, DYS456, DYS558, DYS635, Y-GATA-H4. The size of the YHRD is increasing every day and this has allowed to observe through the Y-STR haplotypes the European population substructure, since population groups do not present significant differences within groups but among groups (Roewer et al. 2005).

Since the first Y chromosme STR was discovered (Roewer et al. 1992), new Y-STRs, 220 approximately, were identified, but those having a greater forensic interest are included in a list compiled by the NIST, from 1997, and making up a database called STRBase, available online (http://www.cstl.nist.gov/biotech/strbase) (Ruitberg et al. 2001).

Recently, the mini Y-STRs utility was observed. Their peculiarity is in the size of the amplicones, that could be generated in a multiplex PCR reaction. Smaller amplified fragments will allow to examine and analyze degraded DNA in a more efficient way (Butler et al. 2003). Asamura (2007) designed the first mini Y-STRs multiplex, including a total of 8 Y-STR loci (DYS522, DYS508, DYS632, DYS556, DYS570, DYS576, DYS504 and DYS540).

1.3.3.2 *Y chromosome disadvantages in forensic analysis*

The lack of recombination is the main Y chromosome disadvantage in forensic. Because of the linkage disequilibrium extention to the great part of the chromosome, Y chromosome is considered as a unique marker, with a loss of informativeness. The information extracted from the haplogroups and haplotypes analysis could not only be referred to a unique individual, but to an extensive masculine lineage. In any case, despite the

informativeness reduction and the possibility to extend the results to a complete lineage, their analysis usefullness cannot be denied, especially in certain cases.

Moreover, because Y chromosome markers are referred to an unilinear transmitted lineage, without modifications from one generation to another, except in occasional mutational events, the haplogroup frequency in a population sample could been drastically altered by past migrations or other recent historical events: this could lead to an incorrect geographic origin allocation.

1.3.4 Y chromosome variability in human populations studies

The complete sequence of the euchromatic region of MSY has been published only recently (Skaletsky et al. 2003), but the analysis of the Y chromosome polymorphisms for population genetic purposes started well before. The discovery of the first polymorphic marker in 1985 (Casanova et al. 1985) was followed by a period in which the search with molecular probes for conventional RFLPs failed to find significant variation in Y chromosome (Jakubiczka et al. 1989; Malaspina et al. 1990). The introduction of the Denaturing High-Performance Liquid Chromatography (DHPLC) increased considerably the number of Y chromosome markers known (Underhill et al. 1997), and since that moment the global Y chromosome tree started to be reconstructed with an increasing resolution. In fact this method has been used to discover more than 200 SNPs and small indels on the MSY (Hammer et al. 2001; Shen et al. 2000; Underhill et al. 2000).

In 2002 the YCC published a paper with two principal aims: to reconstruct a highly resolved tree of MSY and to describe a new nomenclature system (YCC 2002). The association between the binary polymorphisms and the NRY allows us to observe the paternal genetic legacy preserved until present of our species, and to make inferences on the human evolution, population affinities and demographic history.

The constant growth in the number of SNPs and the different nomenclature system used by each author for the Y haplogroups, made the comparison and the data exchange between laboratories very complicated. For that reason the YCC created (http://www.cstl.nist.gov/biotech/strbase) a nomenclature system for the Y chromosome human tree binary haplogrous (figure 1.5). Besides developing a standard nomenclature, in this article authors also tried to show haplogroups equivalences for all the previous

publications with haplogroup names corresponding to the new nomenclature, although in many cases it was impossible to establish a correlation due to the great inconsistencies existing between these first nomenclatures (Capelli et al. 2001; Hammer et al. 2001; Jobling and Tyler-Smith 2000; Karafet et al. 2001; Semino et al. 2000; Su et al. 1999; Underhill et al. 2000) and the one established by the YCC. Until that date the Y chromosome SNPs number described, ascended to 245 defining a total of 153 haplogroups.

The position situated at the phylogenetic tree root was determined comparing homologous NRY regions of closely related species (e.g., chimpanzees, gorilas and orang-outangs), sequenced to determine the ancestral state in human polymorphic sites (Hammer et al. 2001; Underhill et al. 2001). On the tree a few main monophiletic clades can be observed, and a capital letter is assigned to them, starting with the A (it would be first haplogroup in the tree strating from above after the root; figure 1.6) and then in alphabetical order until letter R. Altogether there are 19 letters, also considering Y, the haplogroup including all the others from A to R. At the moment, the phylogenetic tree extended enormously, adding two more clades, S and T (Karafet et al. 2008).

Nowadays, 600 Y chromosome SNPs are approximately described and added to the phylogenetic tree, and so a few readjustments were taken when naming certain haplogroups (Karafet et al. 2008); in addition, due to the great number of branches generated by the new Y-SNPs, the size of the tree also increased considerably. In order to answer to all these nomenclature problems, a new interactive Web page was created: snpreferencedatabase (http://www.snp-y.org/), whose main objective is to support the validation and application of the Y-SNPs, and at the same time being a markers collection for the main population clusters identification.

The contribution of the global Y chromosome tree to the debate on the origin of modern humans soon became clear. In fact, like mtDNA, Y chromosome tree rooted in Africa, with haplogroups A and B restricted to sub-Saharan Africa (Underhill et al. 2001). The haplogroups found in the rest of the world all derive from B sister-clade. This suggested an African origin for modern human populations, and supported the "Out of Africa hypothesis". However, the estimates of the TMRCA were particularly recent when compared to the mtDNA ones (177 kya, (Ingman et al. 2000)), being 46 (16-126) kya using 8 STRs (Pritchard et al. 1999) and 59 (40-140) kya using SNPs (Thomson et al. 2000). The two loci

are in fact supposed to present the same effective population size, the main factor influencing the TMRCA estimates under neutrality. However, the effective population size of Y chromosome has been hypothesized to be lower than mtDNA, due to different possible factors: a higher variance in male reproductive success, natural selection, higher variance in mtDNA mutation rates, stochasticity in the evolutionary process, or questionable assumptions on generation time (Garrigan and Hammer 2006; Jobling and Tyler-Smith 2003). Concerning natural selection, there have been several attempts to apply neutrality tests based on nucleotide diversity to the Y chromosome variation, but it was very difficult to distinguish between signals of population expansion and selection (Jobling and Tyler-Smith 2000, 2003).

Although a few examples of consistent natural selection have been observed, the pattern of Y chromosome haplogroups distribution, compared to other markers, does not seem to have been influenced by selection. For this reason the topology of the tree is still considered reliable (Jobling and Tyler-Smith 2003).

A strong correlation between the Y chromosome variation and geography has been observed in different studies. This property has been used to understand migration patterns, population substructuring and admixture between different populations. Actually, a good Y chromosome lineages representation around the world is available, even though it is possible that some important lineages are still to discover, being necessary a wider sampling in some areas of the planet .

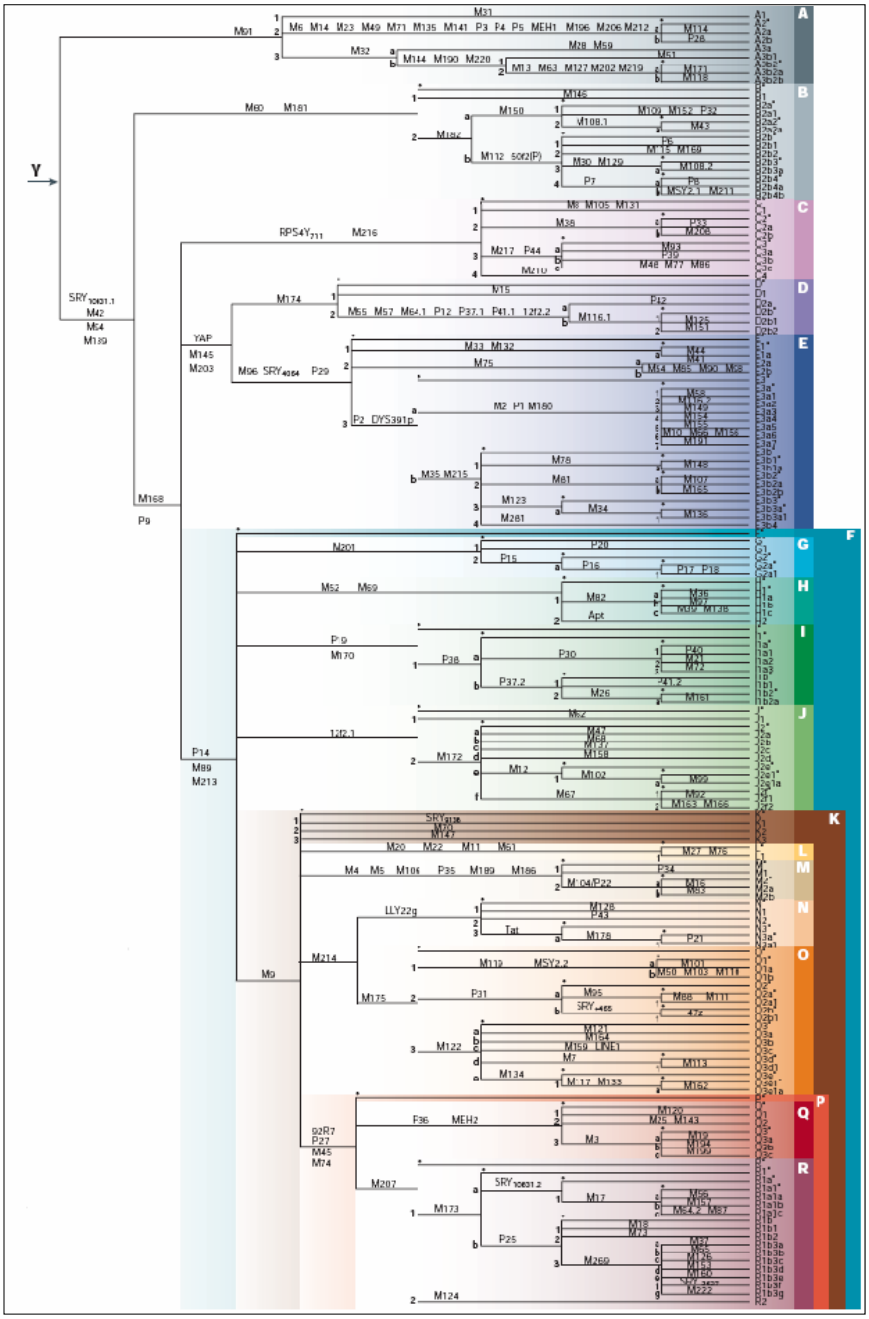


Figure 1.6 Y Chromosome phylogenetic tree (Jobling and Tyler-Smith 2003)

Global Y chromosome haplogroups distribution (figure 1.7), starting from the main clade "A" and in alphabetical order until last main haplogroup "T":

Haplogroup A, in the phylogenetic tree (figure 1.6) corresponding with the oldest branch, appears uniquely in Africa; the same happens to haplogroup B that presents the highest frequencies between hunther-gatheres groups of Ethiopia and Sudan (Cruciani et al. 2006; Semino et al. 2002).

Haplogroup C had a greater dispersion including Center, South and East Asia (Karafet et al. 2001; Kayser et al. 2003; Ke et al. 2001; Underhill et al. 2001; Wells et al. 2001; Zerjal et al. 2003).

Clade D presents the highest frequencies in Tibet (50%) and Japan (35%), but also in other Central and Southeast Asia regions (Cox and Mirazón Lahr 2006; Karafet et al. 2001).

Haplogroup E is one presenting major ramifications, with a great number of described subhaplogroups (Cruciani et al. 2004; Cruciani et al. 2006; Cruciani et al. 2002). Starting with E1a and E2 lineages, described in the North-West of Africa and followed by E1b1, that presents a wide geographic distribution with two of its well defined sublineages: E1b1 in all Africa and in Afro-Americans, and E1b1b1 in West Europe, North Africa and Near East.

Within haplogroup F, smaller clades F*, F1 and F2 appear in the Indian population (Kivisild et al. 2003).

Haplogroup G appears with more frequency in the Caucasian region, but also presents significant frequencies in the Mediterranean and the Middle East areas (Cinnioğlu et al. 2004; Nasidze et al. 2004).

Haplogroup H also appears in India, as it happened with "F", but it has not been deeply studied yet (Kivisild et al. 2003; Sengupta et al. 2006).

Haplogroup I is haplogroup clearly European. It is of most frequent between the Northwest European populations (Rootsi et al. 2004).

The characteristic of haplogroup J, on which there is an almost unanimous agreement, is the dispersion this haplogroup experienced when the Near East individuals West migration took place, carrying this haplogroup to North of Africa, Europe, Central Asia, Pakistan and India (Quintana-Murci et al. 2001; Sengupta et al. 2006). Although other authors, as Di Giacomo (2004), consider this haplogroup the dispersion of the Greek world expansion.

Haplogroup K is the ancestral haplogroup for the "L" to "R" groups (the same happened to haplogroup F, ancestral for "G" to "R" clades). In addition, within K clade, there are some minor clades K*, K1, K2, K3 and K4 distributed all over the world but with low frequencies (Kayser et al. 2003).

Haplogroup L is another main lineage, and it is distributed in India and Pakistan, as well as in the Middle East and, occasionally, in European populations, especially in Mediterranean countries (Kivisild et al. 2003; Sengupta et al. 2006).

Haplogroup M presents its higher frequencies in Melanesia, being its presence limited to the Papuan languages geographic area (Capelli et al. 2001; Cox and Mirazón Lahr 2006; Hurles et al. 2002).

Haplogroup N at the moment presents a wide distribution and probably originated in internal Asia and South of Siberia (Rootsi et al. 2007). The more frequent subclades are N1c, that probably appeared in the region now corresponding to China, expanding from here to Siberia and the East of Europe. The other subclade is N1b, presents in high frequencies in Uralic populations, especially in the Finno-Ugric population.

The O lineage approximately represents 60% of the East of Asian chromosomes; the sublineage O3 has the highest frequency and is absent outside East of Asia, whereas the haplogroups O1 and O2 appear in Malaysia, Vietnam, Indonesia, the South of China, Japan and Korea (Cox and Mirazón Lahr 2006; Hammer et al. 2006; Shi et al. 2005).

A lineage detected in low frequencies in Caucasus and India is haplogroup P (Nasidze et al. 2004; Underhill et al. 2000).

Haplogurpo Q is distributed in Asia, America, Europe and Near East, whereas its subclade Q1a3a is associate almost exclusively with American Native population (Bortolini et al. 2003; Zegura et al. 2004).

Until the recent Karafet publication (2008), the last main clade was corresponding with the amplest haplogroup R. The majority of the individuals belonging to this haplogroup are in the R1 subclade, that is represented mainly by two lineages: R1a and R1b (Brión et al. 2005b; Capelli et al. 2006; Cinnioğlu et al. 2004; Wells et al. 2001). Probably R1b1 haplogroup corresponds to the descendants of the first modern humans who entered Europe. At the moment, it is the more frequent haplogroup in the West of this continent; it also appears in North of Africa and in low frequencies in Iran and Korea; peculiarly it is also in America and Australia but this is probably due to the recent European migrations. On the

other hand, R1a and R1a1 haplogroups are in high frequencies in Central and the West Asia, India and in Slavic population from East of Europe; the distribution of these haplogroups is probably related to the Kurgan expansion. R2 subclade has its higher frequency in South Asia, with low frequencies in Caucasus and Central Asia.

In the previously mentioned Karafet publication (2008), two new main haplogroups are described: S and T.

Clade S, previously called K5 (Scheinfeldt et al. 2006), is mainly distributed in Australia and Indonesia, whereas clade T, previously denominated K2, is observed in low frequencies in Middle East, Africa and Europe.

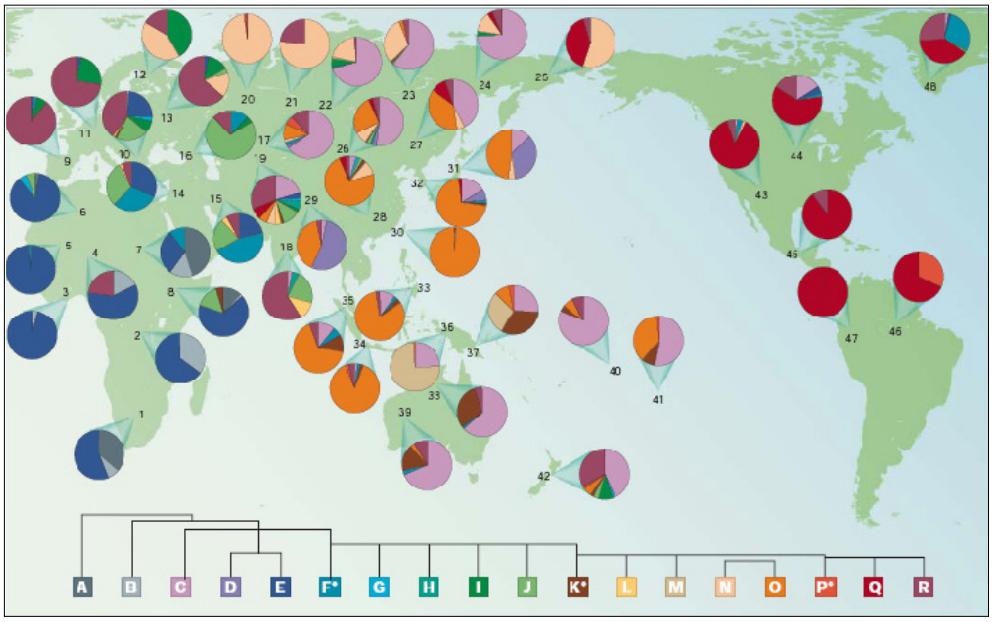


Figure 1.7 The global distribution of Y chromosome haplogroups (Jobling and Tyler-Smith 2003)

The methods used to date the Y chromosome haplogroups phylogenetic order consider the Y STRs diversity within each lineage. The SNPs are rare or unique evolutionary events, that generate male/paternal lineages with which we can study the STRs variability accumulating during the time. Thus, the oldest is a lineage, the greater will be STRs variability.

Besides considering the mutation ages defining the main clades in the NRY haplogroups tree, it also is possible to calculate in a similar way the TMRCA. Few values for the TMRCA were calculated, like in Pritchard work (1999), where using 8 microsatellites valued a TMRCA of 46,000 years (uncertainty of 120,000 years), while Thomson et al (2000) using the SNPs variation of a 64 Kb Y chromosome DNA sequence, obtained an esteem of 59,000 years. All these values of TMRCA are lower than the expected, especially if compared to the values obtained with mtDNA or autosomal loci, and this is partly dued to the reduced population size and to the Y chromosome evolutionary neutrality. As previously mentioned, haplogroups A and B are restricted to Africa, and are distributed along this continent in low frequencies. After few studies, haplogroup A is now considered as part of a proto-African Y chromosomes gene pool (Cruciani et al. 2002). Nowadays the most elevated frequencies of this lineage are in Ethiopians and mainly Khoisans. The second oldest haplogroup is B, that appears in South African Bantu population, central Africa and, for some lineages, Sudan and Mali Pigmys, and moreover, but in lower frequencies than haplogroup A, in Khoisans and Ethiopians. Anthropological data seem to suggest that, during the interglacial period approximately between 130,000-90,000 years, the recent human populations also expanded through Africa, reaching the North and the South Sahara and East Africa, but it seems that the period between 90,000-50,000 is the one associated with the A e B lineages expansion time. Although the fossil and archaeological evidences are too poor to prove this hypothesis, nevertheless it is favored by the paleoclimatic information, which indicates that about 70,000 years ago the glacial climate produced a fragmentation of the African continent in very different and isolated environments between the Northwest, the Northeast and the South. This period between 90,000-50,000 years before the present, could be the responsible for the diversity found outside Africa, as a result of the diversification process produced within the African continent (Underhill et al. 2001).

The first migratory movement outside the continent probably took place around 50,000-40,000 years (Goebel 2007) and overlaps with the entrance in Eurasia of the first modern human, but previously to this event, the Y-SNP M168 (haplogroup CF) mutation generated within Africa. The individuals that entered Eurasia carried this mutation and began to replace the Neanderthals Y chromosome lineages, present in this region until that moment. The dispersion of this lineage started from "the Horn of Africa" towards India and through the Levantine Corridor located in the Middle East (Cavalli-Sforza et al. 1994).

Actually, the sub-Saharian population presents the A, B and E main haplogroups. Haplogroup E is the most frequent and with a high frequency in the Middle East and South of Europe, with occasional presence in Central Asia, Pakistan and America, showing the highest frequencies for the E1b1a subclade in sub-Saharian Africa and for the E1b1b1 subclade in the North and East Africa, in the Mediterranean and Europe (Cruciani et al. 2004; Cruciani et al. 2007). The wide E1b1a lineage distribution is possibly related to the "Bantu agricultural expansion", approximately 3,000 years, with also some Y chromosome replacement. On the other hand, for the E1b1b1 lineage, it probably arised in the "Horn of Africa" region, 24,000-27,000 years ago, expanding from here towards the North and the Northwest of Africa and reaching the Mediterranean. At the end of the Pleistocene it probably dispersed towards the West of Asia, and entering from the Middle East towards the South of Europe, during the agricultural Neolithic expansion (Cruciani et al. 2007; Luis et al. 2004; Semino et al. 2004).

Another mutation, Y-SNP M213, originated in Africa from the populations that already had the M168 mutation, and its dispersion took place from the East of Africa, crossing the Levantine Corridor, towards Eurasia 45,000-30,000 years ago. Paleanthropological evidences also show an expansion of the modern humans from the East, agreeing with the Upper Paleolithic cultures and with the Neanderthals extinction from the Middle East and Europe. The populations carrying this mutation would diversify later in the different clades and subclades, containing F parental haplogroup. Another lineage with a wide dispersion is haplogroup C, that is mainly in Asia, South Austro-Melanesian, Micronesian, Polynesia and North America. This haplogroup arise from the populations that left Africa towards South Asia and carring the mutation M168 (50,000-40,000 years ago). Within this haplogroup, the C3 sublineage is one of most representative, since it is scattered through Turkey, Central and East Asia and, in addition, it is the only C haplogroup variation found in North America.

An exclusively Asian haplogroup is D, uniquely found in Japan and Tibet. This lineage probably had low frequencies, and that facilitated its extinction in the majority of the Asian populations, except Japan and Tibet, where the genetic drift increased its frequency (Underhill et al. 2001). These two haplogroups, C and D, would have been partly replaced by subsequent expansions of individuals with more recent haplogroups.

The individuals carrying the M213 mutation would also have expanded towards Central Asia (not only towards the Caucasus) and, in the course of time, individuals of these populations probably acquired the Y-SNP M9 mutation. The "carriers" of this mutation had to expand widely since they gave origin to several lineages with independent mutations, like haplogroups O, N and P, that distributed in the greater part of Asia moving other previous haplogroups, like C. An example is haplogroup N, that expanded towards West from Asia interior/South Siberia about 12,000-14,000 years ago to arrive to Eastern Europe (Rootsi et al. 2007). Another case is haplogroup M, that arrived to New Guinea and moved partly also haplogroup C.

Within parental haplogroup P, haplogroup R arose in the North of Asia around year 30,000 of our era and within this R1b, that began to disperse towards the West reaching Europe, the Caucasus region, the Middle East, Central Asia and North of India and Pakistan. Another lineage within haplogroup P is Q, that expanded widely through the North Asia steppe, as well as to the center of Asia, North of India and Pakistan. Within this haplogroup Q, the sublineage Q1a3a is characteristic because probably expanded later from Siberia to America.

Different studies realised on the Y chromosome no recombining SNPs, reveal indications of an expansion happened 25,000-20,000 years ago from the East towards the Mediterranean basin and that is related with the haplogroup I presence in Europe. To these expansion events followed a population contraction period associated with the last glacial event of 18,000 years ago and that is also in accord with the anthropological records. The situation radically changes when the climate ameliorated producing R1a and R1a1 haplogroup expansion in Europe from a possible glacial refuge in the Ukraine region approx 10,000-15,000 years ago, as well as the E3b haplogroup Neolithic expansion from the Middle East towards Europe, with also G and J2 haplogroups.

Another great population movements is the one related to haplogroup O, that has high frequencies in the East of Asia and also presents in central Asia and Polynesia. This expansion probably originated in the North of China and its distribution could be related to rice and millet agriculture (Su et al. 2000) about 7,000 years ago, approximately corresponding with the beginning of Polynesia colonization, uninhabited until that moment. Diverse hypotheses could explain the population origine that colonized these islands (Hurles et al. 2002; Su et al. 2000).

Possibly one of the most interesting migration models, and recent too, correspond to the entrance of the first humans in the American continent during the Pleistocene. The main haplogroups founders found seem to be C and Q. Q presents a higher frequency throughout all the continent, whereas haplogroup C frequency is much more reduced. One of the main problems to explain, is the way the Y chromosomes entered the American continent, considering its isolation from the rest. In order to solve that, besides the genetic data we leaned on paleoclimatic data, which indicate that during the Würm or Wisconsin glaciations, started 100,000 years before present and ended approximately 12,000 years ago, a series of glacial phenomena, that implied the oceans level reduction and, in some cases, the appariton of terrestrial connections between several points of the planet. There is an almost total consensus on the entrance in America through Beringia.

1.4 Autosomes

Autosomal polymorphisms, and more precisely the autosomal microsatellites, are the election tools for forensic genetic analysis. DNA profiles based on a set of autosomal microsatellites owe their enormous variability to three processes: (1) mutation, which generates new alleles, (2) independent chromosomal assortment and (3) recombination, which re-assorts the alleles each generation, so that only identical twins share the same genetic profile. The first attempt to reconstruct human evolution with genetic data of living populations was done by Cavalli-Sforza in 1967 (Cavalli-Sforza and Edwards 1967). The genetic distances among 15 populations were calculated, on the basis of the limited data known at that time, using ABO, MN, Rh, Diego and Duffy genes, with a total of 20 alleles. A large amount of data is now available on autosomal DNA polymorphisms (e.g. HLA, GM, Alu), including both biallelic markers and microsatellites. The study of autosomal polymorphisms will certainly be intensively developed in the future, as autosomes represent the major component of the genome.

As we said before, very different polymorphisms distributed throughout the human genome can be classified like:

- Sequence polymorphisms: produced by the change of one (point mutation) or more nucleotides in a DNA sequence. Point mutations are the most frequent variation type in the human genome; every 500 or 1000 DNA bases, a polymorphic base exists.
- Length polymorphisms: produced by insertions or deletions of one or more nucleotides. This type of polymorphism is the more frequently observed in repetitive DNA, like minisatellite and microsatellite DNA, whose polymorphism is due to differences in the repetitive unit number of each allele; in addition, these mini and microsatellites also present combined sequence and length polymorphisms.

Nowadays, in legal medicine, DNA polymorphisms analysis constitutes an established routine work, and the classic markers used are the microsatellites and mtDNA. New tools were added, especially in situations where the classic markers started to fail: autosomal SNPs and mini STRs.

1.4.1 Autosomes variability

1.4.1.1 *STRs*

About the 40% of the human genome is constituted by repetitive DNA. It includes an ample series of sequences distributed throughout the genome, divided in two different typologies: the dispersed repetitive DNA and the tandem repetitive DNA. The repetitive tandem regions have being used in the forensic genetics field practically from the beginnings, because of their fixed position in the genome, their easy detection and high variability. The microsatellites loci are at the moment the election polymorphisms in case studies (Evett and Weir 1998).

STRs mechanisms of variability. The molecular mechanisms originating the number of repetitions variation are essentially two: slipage and the unequal meiotic crossover (Alberts et al. 1994). Slipage is a DNA polymerase copy error in which a same STR repetition nucleus sequence skips or copies twice, adding or removing a repetition, and generating a new allele, i.e. a repetition higher or smaller than the original one. The unequal meiotic crossover is due to a cut and pairing mistake in the chromosomal fragments during the meiosis as a result of the high similarity of the locus sequence; so the cut in the DNA chain is in a different place for each one of the homologous chromosomes.

In the STRs loci, slipage is the main mechanism, verified observing the Y chromosome non-recombining region where the number and the variability of the STRs is not significantly smaller than the rest of chromosomes. The probability of slipage error is greater when the sequence homogeneity is higher, so whenever the repetition unit is smaller, the greater will be the system mutation rate. So, for the dinucleotide STRs (2 bp unit repetitions), although they potentially would have a greater degree of polymorphism, they cannot be used in the forensic field due to their instability

Besides the described processes, the sequence mutation also acts on microsatellite loci. A consequence of this is the presence of intermediate alleles, increasing more the variability of these regions. These intermediate alleles are consequence of a mutational event, a small deletion, that can generate an allele with a number of repetitions plus an incomplete one, giving a new intermediate extension between two preexisting alleles (Wiegand and Klintschar 2002). These intermediate alleles can, in addition, undergo to slipage errors with the same frequency of the typical ones so that, from them, it is possible to generate new

group of alleles, increasing even more the variability of the system, like for example the FGA system (Mills et al. 1992).

1.4.1.2 *Autosomal SNPs*

At the moment, because of the automatization and miniaturization of the DNA molecular variability detection methods, binary polymorphisms, and especially SNPs, are taking a very important paper in the forensic field.

The main problem to consider at the time of using binary polymorphisms, is their low level of variability, so that the first question is about how many SNPs should be used in order to obtain, for example, the same resolution of 17 STRs loci (Fung et al. 2002; Gill 2001).

This type of polymorphisms, because of their biallelic state and low mutation rate, probably have a population substructure more important than the microsatellite loci, so that severe validation studies before their application in forensic and paternity cases are necessary (Chakraborty et al. 1999). One of the main objectives in implementing the use of these polymorphisms in the forensic routine, is the development of robust multiplexes, that will make possible the simultaneous amplification of 20 or 30 SNPs in a single reaction, since the amount of DNA in forensic samples often constitutes a problem. In this sense, the development of new technologies is supposing an important advance, allowing to analyse in a short time and from small amounts of degraded DNA, a great number of polymorphisms.

1.4.2 Autosomes applications in forensic genetics

1.4.2.1 *STRs in forensic genetics*

The fact that these markers are in non-selective zones of the genome, makes possible the multiple alleles conservation in the population gene pool, generated by slipage and unequal meiotic crossover, and not eliminated by negative selection; in addition, the high number of alleles and the fact that none of them has effect on the individuals fitness (not being favored nor underprivileged by natural selection) permit their presence with relatively low frequencies in the population. That is the reason why STRs have a high informativeness for their use in identification or connection between individuals.

The relatively short STRs extension makes their analysis possible with PCR techniques, allowing a simple typing, extremely sensitive (being able to work successfully with picograms amounts of DNA) with degraded samples (typical in the forensic field), and also in multiplex reactions (Evett and Weir 1998).

STRs typing. Actually, few STRs multiplex amplification kits exist on the market, perfectly optimized for the forensic work, designed to operate with automatic sequencers, using a minimum sample quantity and with a great genotyping accuracy. These kits incorporate a fluorochrome to the amplification primers, so that they can be detected by the sequencer, giving an exact measurement of the electrophoretic mobility of the fragments, on the basis of the time they run through all the capillary extension. The high polymorphism degree, and therefore the high informativenees of these markers, along with the short length of these loci, have been the reasons why STRs are the first election polymorphisms in forensic work.

Disadvantages. In spite of their undeniable advantages, STRs continue presenting a minimum 90 bp amplicon length, which is the main disadvantage of these systems compare to SNPs. In conditions of highly degraded DNA, STRs can begin to fail especially those of higher molecular weight. Generally, the greater is the DNA degradation, the smaller will be the probability to find the DNA fragments with a high molecular weight, so that heaviest STRs will begin to fail. Long fragments are difficult to detect, and almost impossible in extreme conditions.

Artefacts amplification. For the STRs used in forensic work, when the reaction efficiency is high, it is common to observe shorter repetition units, also called stutter bands. Nevertheless, these artefacts are easily identifiable in a normal reaction because the quantity is smaller than the true reaction product, so that, generally, they do not bring to any result misinterpretation. But in criminal cases, the stutter bands can be an error source, since it is possible that the DNA sample to analyze is composed by material from different donors and in different proportions. The lower material profile will appear, then, like accompanying bands of smaller quantity than those of the main profile. In this case, the stutter bands would be indistinguishable from the minority profile, making difficult the correct data interpretation. The greater STRs mutation rate of the shorter nucleus sequence, can make these artefacts masking the true reaction product (Evett and Weir 1998).

A second type of extremely common artefacts are the fluorescence residuals, or crossed bands. These artefacts are dued to the fluorochromes marking the PCR products: in

fact, they do not beam exclusively in a unique wavelength, but in a series of light spectrum frequencies. Nevertheless, if the amount of molecule, and therefore of energy emitted by the fluorescent marking, is sufficiently high, the sequencer detector could detect this emission like a peak in a color band that does not correspond to the labelled molecule, masking in that way an underlying signal of another marker, or altering the correct electopherogram reading.

miniSTRs. In response to the high weight STRs amplification problems, mini STRs typing has being introduced in forensic (Butler et al. 2003; Coble and Butler 2005; Grubweisser et al. 2006; Opel et al. 2006). This new forensic STRs typing approach is based on the use of shorter microsatellites and the redistribution of these markers in a multiplex, placing the amplification primers closer to the repetitive region, and so reducing the amplicon maximum length. In the mini STRs commercial multiplexes, recently designed for forensic application, with nine markers including amelogenine (informative marker on sample sex), the amplicon sizes oscillates between 90 and 300 bp (AB).

In spite of the amplicons reduction, the problem persists in more degraded samples, since it is impossible the simultaneous typing of a great number of markers maintaining the amplicon size below 120 bp for all of them.

1.4.2.2 *Autosomal SNPs*

These genetic markers have been used for identification inferences and kinship relations in a great number of fields, like anthropology, ecology and forensic sciences.

This type of polymorphisms are usually called biallelic because in the majority of the cases they present only two possible variants, one ancestral and one derived. In spite of this, there are in the genome gene pool numerous nucleotide positions with three and up to four possible variants (Phillips 2005).

Unlike STRs, SNPs have not been used for identification inferences, since, being the majority of them biallelic, they have a lower resolution power than microsatellites. This disadvantage can be simply avoided increasing the number of analysed markers, until the resolution power obtained is sufficiently elevated, which is relatively simple given the unique characteristics of the SNPs. The analysis of a relatively reduced number of SNPs (about 50 loci) constitutes a sufficiently effective tool for identification and paternity tests, even

between close relatives (Anderson and Garza 2006; Gill 2001). SNPs selection should be in accord with three characteristics: 1) the independent segregation between the markers (in order to avoid the trasmission in block and have a complete indipendent information), 2) the degree of polymorphism (evaluated according the allelic frequency of the minority variant in the population; a reduction of the variability within certain limits does not affect the exclusion probability and if the number of markers used is sufficient, the loss of variability will be like the awaited one) and 3) the quality of the flanking sequence.

Autosomal SNPs application in forensics. Lately, SNPs typing has become a frequent tool support in forensic genetics laboratories. The use of these markers, as support to the traditional ones, is possible in three different cases: the low efficiency of the traditional markers with highly degraded DNA, to obtain information about individuals in absence of samples for profiles comparisons and in case of deficiency of the classic markers in a paternity case, in which the possible ancestor is a close relative.

Application to degraded samples. In case of extremely degraded DNA, SNPs give better results than the microsatellites analysis because they work with shorter amplicones and are more resistant to the PCR inhibitors action.

The use of SNPs in these cases is like a support tool. Working with degraded DNA still permits to obtain, in the majority of the cases (corpse identification, study of biological vestiges, work with histological samples including in paraffin,…) a STRs profile, even if partial. In these cases, the addition of the information provided by a high number of SNPs, which have a higher rate of success in highly degraded samples, extends the information, reaching values that widely pass the acceptable minimums for a correct identification.

The exponential double strand DNA amplification with multiple pairs of primers in a unique reaction constitutes an efficient use of the DNA available. On the other hand the design of these reactions allow to work with short amplicones (between 60 and 120 bp), that is compatible with fragmented molecules, typical of degraded DNA (frequently smaller of 150 bp) so that the probability of an allelic dropout is minimum (Sánchez and Endicott 2006).

Application to the prediction of the most likely population origin. The forensic genetics aim in legal medicine, is mainly to determine the relations between the DNA profiles obtained from the biological samples related to every case.

The typing of SNPs located in non-selective DNA regions allow to obtain certain informations about the individual, like for example, its possible geographic origin. For that aim, the most interesting SNPs property is their reduced mutation rate, $2x10^{-8}$ per generation, so that the probability a same change takes place twice independently is so low that this will not occur. It is possible, then, that when a change appears between the individuals of a population, because of the reduced gene flow between the same populations and the other ones, this change is restricted to the origin population and to the ones originated from it. Because of the completely random behavior of the evolutionary forces affecting exclusively the molecular evolution of the non-functional DNA regions, the variant frequencies in the gene pool of different populations are possibly drastically different. These differences due to the random evolutionary factors would allow to establish, on the basis of the variant in the studied profile, the more probable population origin of the sample. The markers selected for this purpose in the last years, have been the mtDNA and the Y chromosome (Santos and Tyler-Smith 1996; Torroni et al. 1996).

In Phillips et al study (2007), a multiplex of SNPs located in the non-functional autosomal zones was developed, allowing the estimation of an individual population origin probability with great accuracy and security. Thirty-four neutral SNPs segregating indipendently were used, with a high intergroup variance. Considering the frequency distribution of these SNPs in two given populations, A and B, for a SNP with alleles 1 and 2, the allelic frequency of both variants would be extreme in one of the populations (0.9-0.1) and balanced in the other (0.5-0.5). These differences between populations in the allelic frequency for each one of the 34 SNPs, allow, given a DNA profile, to give an esteem of the probability that a profile originated in a given population. The 34 SNPs were selected in order that their frequencies were determining one of the three great human populations groups: african, caucasian and asian. Studying a high number of independent markers, the disadvantage presented by the unique locus information of mtDNA and Y chromosome disappears. Working with autosomal, for the fact they recombine, do not give any lineage information, but the probable gene pool in which the profile could have been originated.

1.5 The peopling of Africa

The spread of the first communities of modern *Homo sapiens* within the African continent, after their putative origin in East Africa 150,000-200,000 years ago, is a complex topic which still needs to be elucidated. The limited availability of the fossil record for modern *Homo sapiens* in the continent is one of the most limiting factors to the study of this subject. On the other hand, more recent population processes that shaped the peopling of sub-Saharan Africa have been studied in greater detail. Within this framework, Sub-Saharan Africa deserves special attention since it is one of the areas that more than others suffers from the lack of fossil record, it was intensively influenced by the recent expansion of Bantu languages, and it is currently inhabited by one of the last hunter-gatherers communities in Africa, the Pygmies.

Africa is an important region to study human genetic diversity because of its complex population history and the dramatic variation in climate, diet, and exposure to infectious diseases, which results in high levels of genetic and phenotypic variation in African populations. A better understanding of levels and patterns of variation in African genomes, together with phenotype data on variable traits, will be critical for reconstructing modern human origins and the genetic basis of adaptation to diverse environments (Campbell and Tishkoff 2008).

Africa is a region of considerable genetic, linguistic, cultural, and phenotypic diversity. There are more than 2,000 distinct ethno-linguistic groups in Africa, speaking languages that constitute nearly a third of the world's languages (http://www.ethnologue.com/) (Campbell and Tishkoff 2008).

The pattern of genetic variation in modern African populations is influenced by demographic history (e.g., changes in population size, short and long range migration events, and admixture) as well as locus-specific forces such as natural selection, recombination, and mutation. For example, the migration of agricultural Bantu speakers from West Africa throughout sub-Saharan Africa within the past ~4,000 years and subsequent admixture with indigenous populations has had a major impact on patterns of variation in modern African populations (Pilkington et al. 2008; Quintana-Murci et al. 2008; Reed and Tishkoff 2006; Tishkoff et al. 2007; Wood et al. 2005). Although Africa is critical

for understanding modern human origins and genetic risk factors for diseases, it has been under-represented in human genetic studies. Much of what we currently know about genetic diversity is from a limited number of the ~2,000 ethno-linguistic groups in Africa, and the majority of these data are from mtDNA and Y chromosome studies. Large-scale autosomal studies of African genetic diversity are only now beginning to become available. The study of variation in extant populations can provide novel insights to the general picture of the ancient peopling of the area. However, caution should be used because recent events could have affected the genetic composition of the populations inhabiting the area.

1.5.1 mtDNA variation in Africa

One of the pioneer studies of mtDNA variation in human populations produced a tree that showed a deep split between sub-Saharan Africans and non-Africans with a coalescence dating back to 200,000 years ago (Cann et al. 1987). This study paved the way for further investigations into human populations, of which sub-Saharan African populations were considered to be of particular interest and importance because of their essential role in any genetic test of the hypotheses concerning the emergence of modern humans (Harpending et al. 1993). It was observed early on, that most sub-Saharan mtDNAs (from 70% to 100%, depending on the population considered) present a specific *Hpa*I restriction site at position 3592 (Torroni et al. 1994). These haplotypes were subsequently assigned to a lineage which was conventionally termed L (Chen et al. 1995; Salas et al. 2002) and which contains several super-haplogroups (Salas et al. 2004), with relative haplogroups.

The accumulation of studies on L lineages in recent years led to a considerable revision of their reciprocal relations, and of their nomenclature. Haplogroup L0 includes four sub-haplogroups: L0a, L0d, L0f and L0k. All of these are mainly spread in South and East Africa. L0d and L0k are found almost exclusively in southern Africa Khoisan-speaking populations, with the exception of few L0d lineages found in Khoisan-speaking populations from Tanzania (Gonder et al. 2007; Tishkoff et al. 2007). L0a is common in eastern, central and south-eastern Africa, while it is almost absent in northern, western and southern Africa. An origin of this clade in East Africa seems likely (Salas et al. 2002). Finally, L0f is rare and appears to be geographically confined to East Africa (Salas et al. 2002).

Haplogroup L1 is composed of L1b and L1c, both of which are almost absent in East Africa. L1b is mainly confined to western Africa, with some overflow into Central and North Africa. It is also common among African Americans. Salas and colleagues proposed a Central African origin for this haplogroup (Salas et al. 2002). L1c occurs at highest frequencies in Central Africa, whereas it is less common in North, West and East Africa. It is also found in African Americans, and reaches its highest frequencies in Western Pygmy populations (Alves-Silva et al. 2000; Batini et al. 2007; Destro-Bisol et al. 2004; Salas et al. 2002). It has been recently proposed that L1c could have originated in ancient times in Central Africa, prior to the separation between the ancestors of present day Bantu-speaking and Pygmy populations (Batini et al. 2007; Quintana-Murci et al. 2008).

Haplogroup L2 includes four sub-haplogroups: L2a, L2b, L2c and L2d. L2a is widespread all over Africa, including Eastern Pygmies (Salas et al. 2002; Torroni et al. 2001). L2b and L2c are common in western Africa, with some overflow in eastern Africa (Salas et al. 2002; Tishkoff et al. 2007; Torroni et al. 2001). L2d is rare and it is mostly confined at low frequencies in western Africa (Torroni et al. 2001).

Haplogroup L3 is composed of several sub-haplogroups. L3b is almost restricted to West and North Africa, and to African Americans (Salas et al. 2002). Its sister clade L3d is also mainly West African and African American, with few types observed in south-eastern Africa (Salas et al. 2002). L3e is widespread all over Africa, with one lineage more frequent in south-eastern Africa (L3e1), and others more common in West and Central Africa (L3e2, L3e3 and L3e4)(Salas et al. 2002). This haplogroup seems to be the one most affected by the Bantu expansion (Salas et al. 2002). L3f seems to have originated in Central and West Africa, but its spread zone is mostly East Africa (Kivisild et al. 2004). L3h was first reported at a moderate frequency in Guinea-Bissau (Rosa et al. 2004) and it has been observed at low frequencies in East Africa (Kivisild et al. 2004). L3i, L3x and L3w have been described recently, and have been observed mainly in north-eastern Africa (Kivisild et al. 2004).

Haplogroup L4 includes L4a and L4g (previously L3g), both showing an East African distribution, with some overflow in North Africa (Kivisild et al. 2004).

Haplogroup L5, previously referred to as L1e (Kivisild et al. 2004), has been observed at low frequencies only in eastern Africa, in Egypt and among Mbuti Pygmies (Salas et al. 2002; Stevanovitch et al. 2004).

Haplogroup L6 has been recently defined and has been observed only in Ethiopian and Yemeni samples. An East African origin for this clade has been proposed (Kivisild et al. 2004).

The distribution of the main L clades within the different macro-areas in sub-Saharan Africa is shown in figure 1.8.

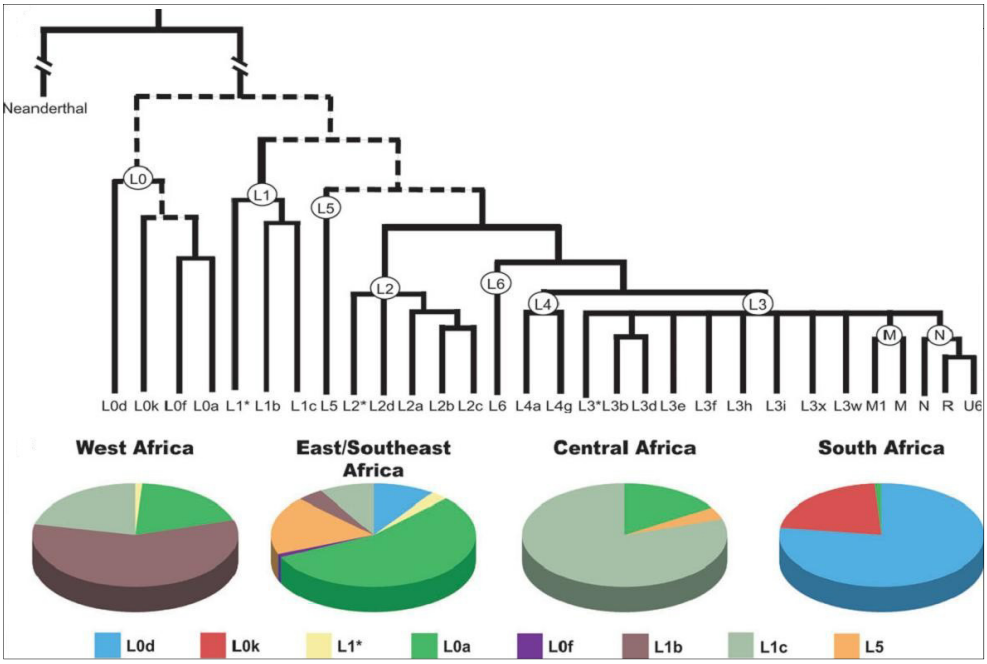


Figure 1.8 Schematic representation of L phylogeny (Gonder et al. 2007)

The prevalence of L0, one of the first branches of the complete phylogeny, in East and South Africa indicates once again these areas as the most probable place of origin of *Homo sapiens*. Furthermore, it has been recently proposed, on the basis of phylogeographic evidence, that during the first phase after their origin, the populations of *Homo sapiens* were deeply separated in different groups (Behar et al. 2008). This finding could explain the today high frequency of specific lineages in specific hunter-gatherer groups.

1.5.2 Y-chromosome variation in Africa

Sub-Saharan African populations are characterized by the presence of four main haplogroups: A, B, E and R (figure 1.6).

Haplogroups A and B are the deepest branches of the Y-chromosome phylogeny and show a wide distribution but are generally present at low frequencies. These are the only two clades to be almost completely confined to sub-Saharan Africa.

Haplogroup A has been found in several Khoisan-speaking populations from South Africa (Cruciani et al. 2002; Underhill et al. 2001; Wood et al. 2005), Fulbe from West-Central Africa (Cruciani et al. 2002), and Sudanese, Ethiopians and Sandawe from East Africa (Cruciani et al. 2002; Tishkoff et al. 2007; Underhill et al. 2001). Outside Africa it was detected in one Sardinian individual, probably due to recent gene flow (Underhill et al. 2001).

Haplogroup B has been found in most of the populations cited above as well as in Biaka and Mbuti Pygmies (Underhill et al. 2001), in several populations of different language groups in East Africa (Knight et al. 2003; Luis et al. 2004; Tishkoff et al. 2007), and in Yoruba, Mossi and Bamileke from West-Central Africa (Cruciani et al. 2002; Tishkoff et al. 2007).

Haplogroup E is the most common and widespread haplogroup in Africa. Two of its subclades (E1 and E2) are quite rare and are mainly present in eastern Africa (Luis et al. 2004). The other clade, E3, is widespread and found at very high frequencies. It is subdivided in two subgroups, E3a and E3b, that show different distributions. E3a has been proposed as a marker of the Bantu expansion (Underhill et al. 2001), and it is present in almost all sub-Saharan Africa, including hunter-gatherer populations from both Central and South Africa (Beleza et al. 2005; Cruciani et al. 2002; Tishkoff et al. 2007; Underhill et al. 2001). E3b has been shown recently to encompass groups with very different evolutionary histories. In Africa it is limited to the East and the North, but it has also been observed in Europe and western Asia (Cruciani et al. 2004; Semino et al. 2000; Tishkoff et al. 2007; Underhill et al. 2001; Wood et al. 2005).

Haplogroup R is actually rare in Africa and is found mainly in Asia and Europe and it is thought to have originated in Asia (Underhill et al. 2001). However, one of its sub-clades, R1, has been found at high frequencies, up to 95%, in some populations from North

Cameroon (Cruciani et al. 2002). The authors proposed a back migration from Asia to sub-Saharan Africa, possibly via northern Africa, to explain this phenomenon. The absence of an evidently similar pattern for mtDNA leaves this topic still controversial (Coia et al. 2005).

1.5.3 Insight Sub-Saharan Africa

1.5.3.1 *Cameroon*

The area occupied by the present-day Republic of Cameroon is of particular importance for bio-anthropological studies, since it may be regarded as a sort of transect that comprises an important part of the vast biological and cultural diversity of sub-Saharan Africa (Campbell and Tishkoff 2008). In fact, it is inhabited by a large wealth of populations, which differ substantially in subsistence strategies, language, social structure, and religion. A primary distinction can be made using a geographic criterion. The populations living in the northern part of the country (provinces *Extrème Nord* and *Nord*), where the Savannah and Sahel habitat predominates, are often referred to as "Sudanese" populations. Their typical subsistence economy (the so called "Sudanic complex") is mainly based on the cultivation of cereals such as kaffir corn and millet (*Sorghum* and *Panicum miliaceum L.*) and the breeding of cattle and sheep (Ehret 1984; Harris 1976). The Sudanese are linguistically heterogeneous, speaking languages belonging to the Afro-Asiatic, Nilo-Saharan and Niger-Kordofan phyla (Greenberg 1963). Some of these populations, the so-called *Montagnards* (Podokwo, Mada, and Uldeme), are thought to be in continuity with the groups that created the oldest nucleus of settlements in northern Cameroon, the "Sao civilization" (Lebeuf 1981; McIntosh and McIntosh 1983). The northern region was first populated around 8,000 ybp. by groups belonging to the original Afro-Asiatic or Hamito-Semitic linguistic stock from the Near East (David 1981). Subsequently, another migration of people from the Sahelian area to the South reached the region of the Adamawa plateau around 4,000 ybp, a likely consequence of the desertification of the Sahara (5,000–7,000 ybp) (David 1981). Finally, the area was recently populated (18th century) by the Fulbe from Nigeria (Mohammadou 1973).

The climate of the southern land (provinces Ouest, Littoral, and Centre) is equatorial and most of the territory is occupied by tropical and equatorial forest. Some of the populations settled in this area are thought to be descendants of the proto-Bantu nucleus that originated on the Nigerian-Cameroon plateau around 4,000 ybp (Bakaka and Bassa),

whereas others (Bamileke and Ewondo) settled more recently (18th century) in the area and adopted a Bantu language (Ehret 1984; Spedini et al. 1999). Their traditional agricultural techniques, the so-called "vegecultural complex" (Harris 1976), use tubers including yam and manioc, while the diet is complemented by the breeding of small animals and poultry.

In contrast to north Cameroon, populations from south are linguistically homogeneous. In fact, they speak languages of the sub-branch Benuè-Congo of the Niger-Kordofan phylum (Greenberg 1963), generally referred to as Bantu.

Given this complex background, Cameroon provides a unique opportunity to study how biological, geographic, and cultural factors interact in determining variation within and among human populations.

The genetic variation of Cameroon populations has been analyzed at protein coding loci (Spedini et al. 1999), mitochondrial DNA (Cerný et al. 2004; Cerný et al. 2007; Coia et al. 2005; Destro-Bisol et al. 2004), and Y-chromosome (Caglià et al. 2003; Coia et al. 2004; Cruciani et al. 2002; Wood et al. 2005). However, there is limited data regarding microsatellite polymorphisms, which are particularly useful for population genetic studies due to their high level of variation (Coia et al. 2009).

1.5.3.2 *Western and Southern Africa*

Recent population processes that shaped the peopling of sub-Saharan Africa have been studied in greater detail in the last years. One of the best known event is the expansion of Bantu languages, which was linked at some stage to agricultural and metallurgical innovations, beginning between 5,000 and 3,000 years ago from the area between Nigeria and Cameroon and involving most of sub-Saharan Africa.

Sub-Saharan Africa is one of the areas that most attracted research interest since the beginning of molecular studies on human populations, because of its putative fundamental role in the first phases of the evolution of *Homo sapiens*, that seems to have originated in the eastern part of this continent 200,000 years ago. The spread of the first communities of *Homo sapiens* from this area to the rest of Africa is a complex topic that still needs to be disentangled. In this context, Central Africa deserves special attention since this is one of the areas inhabited by one of Africa's last hunter-gatherer communities, the Pygmies. These populations are supposed to be in genetic continuity with the first inhabitants of the area,

and therefore the analysis of their variation could provide interesting information on the ancient peopling processes. Furthermore, Central Africa presents several characteristics that makes it an interesting case study in which the application of a molecular approach is particularly fruitful. In fact, from a paleoanthropological perspective, the area under study suffers more than others the lack of fossil record. The earliest modern humans found, date, in fact, at maximum 20,000 years ago. The archaeological and paleoclimatological records have started to be intensively studied only in recent years. These are strongly affected by postdepositional disturbance, which makes their interpretation difficult, leaving the overall picture unstable. Finally, the linguistic record has been considerably influenced by the expansion of Bantu languages, hiding all traces of the languages previously spoken in the area. Therefore, the analysis of genetic variation of populations today inhabiting Central Africa could offer an additional independent approach to the reconstruction of the peopling of the area.

A considerable influence of socio-cultural factors can be hypothesized in the shaping of the diversity of these populations, but also recent events, such as the expansion of Bantu speakers, seem to have generated an homogenizing effect for Y-chromosome variation, that did not show substantial differences among all populations.

Whereas the dissection of single Y-chromosomal clades or sub-clades proved to be useful to shed light on the relations between specific populations/groups and helped reconstruct the demographic impact of migratory and cultural events, a wider and exhaustive phylogeographic analysis may provide indications on areas of the African continent where the extant human Y chromosome diversity first originated. The haplogroups A and B are ideal candidates for this task, given their distribution in Africa and the fact that they represent the earliest lineages to branch off within the Y chromosome genealogy. A detailed phylogeographic dissection of haplogroups A and B in a broad data set of sub-Saharan populations, aims to provide new insights into the complex and poorly investigated dynamics that characterize the preagricultural history of sub-Saharan Africa, with special attention given to the relationships among Pygmy and Khoisan-speaking populations from southern Africa (Batini et al. 2011).

However, the current absence of significant palaeo-anthropological investigation couples with the different possibility of fossil preservation in central Africa and makes the extremely long human fossil record in eastern Africa not conclusive in solving this issue

(Batini et al. 2011). The screening of Y-chromosomal variation at the same level of resolution in additional populations from these regions as well as the analysis of genomic data, is expected to provide further details on the early steps of *Homo sapiens* in Africa (Batini et al. 2011).

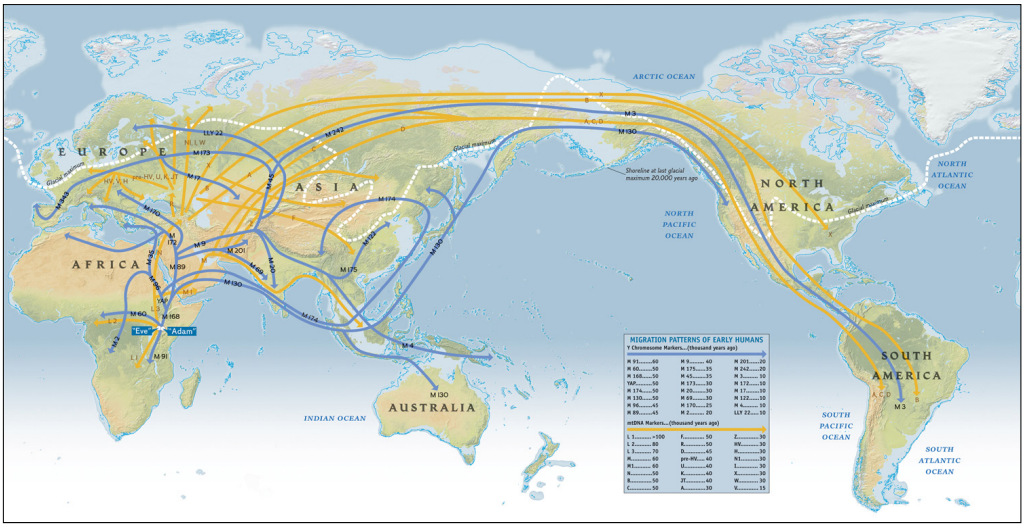


Figure 1.9 Human migration routes beginning about 100,000 years ago, based on mitochondrial (yellow) and Y-chromosome(blue) DNA evidence collected by the National Geographic Society's Genographic Project and other sources (http://www.nationalgeographic.com/).

1.6 The peopling of Europe

Since the first attempts to use biological variation in humans to aid our understanding of early human migrations, the peopling of Europe has been a major research focus (Busby et al. 2011; Cavalli-Sforza et al. 1994; Menozzi et al. 1978). Anatomically modern humans initially entered Europe around 40,000 years ago (Diamond and Bellwood 2003), later retreating to glacial refugia following a worsening climate during the Last Glacial Maximum (25,000-18,000 years ago) (Blockely and Pinhasi 2011). From about 14,000 years ago, Europe was then recolonised from these regions as a result of an improving environment (Blockely and Pinhasi 2011). Some glacial refuge zones of southern Europe (Franco-Cantabria, Balkans, and Ukraine) were the major genetic sources for the human recolonization of the continent at the beginning of the Holocene. Intriguingly, there is no genetic evidence that the refuge area located in the Italian Peninsula contributed to this process (Pala et al. 2009).

The same conditions favoured the development of agricultural technology in the Fertile Crescent some 10,000 years ago (Gamble et al. 2005). Spreading from the Near East westward into Europe, this technology caused a major cultural transition from itinerant hunter-gathering, to sedentary farming which, importantly, created an increase in the rate of population growth (Collard et al. 2010; Gamble et al. 2005), in what has become known as the Neolithic transition (Cunliffe 2001; Jobling et al. 2004). Within this archaeological framework, the debate rages about the relative contributions to modern European populations of the first people of Europe and those who migrated into it with the Neolithic transition, both in terms of their genetic legacy, and as to the processes of migration and succession (Battaglia et al. 2009; Capelli et al. 2003; Capelli et al. 2006; Chikhi et al. 2002; Gallagher et al. 2009; Rowley-Conwy 2009). The true scenario is undoubtedly multi-faceted and complex. Both early works on "classical markers" and more recent studies using the Y chromosome have shown that in Europe genetic variation is distributed along a southeast-northwest gradient. Such observations have been suggested to support a model of demic diffusion for the Neolithic transition in Europe, i.e. that the spread of agriculture also involved an associated movement of people from the Near East (Cavalli-Sforza et al. 1994; Novembre et al. 2008; Rosser et al. 2000; Semino et al. 2000). However, a more recent work, based on different Y chromosome haplogroups in Balkan and eastern European populations,

suggests a role for the cultural diffusion of agriculture from the southeast Europe (Battaglia et al. 2009).

Phylogeographic analysis of European mtDNA lineages suggests that the resettlement of much of Europe occurred through the Mesolithic expansion from the putative Franco/Cantabrian Iberian and Eastern European glacial refugia (Malyarchuk et al. 2008; Soares et al. 2010). Undoubtedly Neolithic Near East expansions made their mark in the current mtDNA gene pool, but founder analysis suggests that their effect was small (around 15%) (Soares et al. 2010). Although based on small sample sizes, analysis of the ancient DNA of Neolithic farmers further suggests a lack of genetic continuity between the farmers and contemporary Europeans (Haak et al. 2005) pointing towards a mostly pre-Neolithic contribution. However, more fresh analysis complicate this picture by finding an additional lack of genetic continuity between ancient hunter-gatherers, as well as Neolithic farmers, and modern Europeans (Bramanti et al. 2009; Haak et al. 2010).

If the Out-of-Africa Hypothesis is correct, then *Homo sapiens sapiens* spread from the African tropics across the Sahara Desert into the Mediterranean region before the last glaciation caused the Sahara dry up. There is at present no archaeological evidence for this apparently rapid population movement. The earliest occurrence of modern humans in the Near East dates to about 100,000 years ago at Qafzeh Cave in Israel. But important fossil discoveries at Mount Carmel and elsewhere tell us that for 50,000 years, both anatomically modern and more archaic, Neanderthal-like humans lived alongside one another in this region, apparently still using old-fashioned, simple tool kits. It is only after about 45,000 years ago that the characteristic blade technology and more specialized artifacts associated with modern humans appear in the Near East, perhaps a response to drier conditions that required more efficient stone tool technology. No one has been able to explain the apparent contemporaneity of archaic and modern humans for such a long period of time within such a small area. Some believe modern humans may have evolved not in Africa, but in the Near East. Others argue the two groups lived in the same area, but had different lifeways and territories. Many more fossil discoveries will be needed to resolve the issue.

Forty-five thousand years ago, Europe and Eurasia were intensely cold, with long, subzero winters. These severe climatic conditions may have inhibited the spread of modern humans into northern areas and onto open plains and steppe-tundra landscapes until the development of more effective, specialized tool kits that allowed the working of bone and the

making of tailored clothing for bitterly cold conditions. Whatever the cause, no anatomically modern humans appeared in Europe before about 43,000 years ago, when, apparently, they crossed the then-dry Bosporus into the Balkans.

The first fully modern Europeans are known to biological anthropologists as the Cro-Magnons, named after a rock shelter near the village of Les Eyzies in southwestern France. They are indistinguishable from us, strongly built, large-headed people whose appearance contrasts dramatically with their Neanderthal predecessors. The anatomically modern ancestors of the Cro-Magnons had settled in southeast and central Europe by at least 43,000 years ago, apparently near Neanderthal groups. Some of them had penetrated into the sheltered, deep river valleys of southwestern France by 40,000 to 35,000 ybp. There they seem to have lived alongside the Neanderthals, but the relationship between the two groups is still little understood, despite some borrowing of tool kits. But by 30,000 years ago, the Neanderthals had vanished and the density of Cro-Magnon settlement intensified considerably. The ancestors of the Cro-Magnons had entered Europe during a brief period of more temperate climate. Even then, climatic conditions and seasonal contrasts may have been such as to require new artifacts and much more sophisticated hunting and foraging skills. These adaptations developed rapidly, indeed spectacularly, after 30,000 years ago, when the climate grew colder. It was during these millennia that *Homo sapiens* finally mastered winter, for it was in northern latitudes that human ingenuity and endurance were tested to the full. The highly successful Cro-Magnon cultures survived from at least 32,000 years ago to the end of the Ice Age, when the glaciers finally melted and dense forest spread over the open plains and deep valleys of central and western Europe.

The open steppe-tundra plains that stretched from the Atlantic to Siberia were a far harsher environment. To live there permanently, Late Ice Age people had to find sheltered winter base camps, have the technology to make tailored, layered clothing with needle and thongs, and the ability to build substantial dwellings in a treeless environment. Only a handful of big-game hunting groups lived in the shallow valleys that dissected these plains before the glacial maximum 18,000 years ago. Thereafter, the human population rose comparatively rapidly, each group centered on a river valley where game was most plentiful, and where plant foods and fish could be found during the short summers. It was here that the most elaborate winter base camps lay, settlements like Mezhirich in Ukraine, which is famous for its finely made mammoth bone framed houses. Some of these groups traded with

neighboring bands over hundreds of miles, exchanging sea shells and exotic tool-making stone with one another.

1.6.1 Insight the Italian Peninsula

Due to the geo-morphological characteristics of Italy, it was one of the favorite destinations for human groups coming from Africa, Middle East and other European locations.

Hominid presence in the Italian peninsula has been complex and extended in time. *Homo sapiens* probably made his first appearance in this area around 30,000-40,000 ybp (Cunliffe 2001). Around 11,000 ybp in the Fertile Crescent new resources became available to humans in the means of domesticated crops and animals. The new technology was now able to support large communities and provided the resources for a demographic expansion (Cunliffe 2001). Technology spread quite fast across the European peninsula, reaching the western fringes just 4,000 years later (Ammerman and Cavalli-Sforza 1984). The related demographic impact is still a matter of debate, but a consensus seems to have been reached on substantial Neolithic contribution in the Mediterranean area (Chikhi et al. 2002; Semino et al. 2000; Simoni et al. 2000). In Italy, Apulia, Calabria and Eastern Sicily were involved since the very beginning in this process as testified by first Neolithic archaeological remains dating around 9,000 ybp. Farming technology appeared in Central Italy, on both sides of Apennines, and in the North East, in the Po and Adige Valleys, only 1,000 years later. In the remaining areas, i.e. North and Central West Italy, farming technology arrived later, around 6,500 ybp and was characterised by a marked continuity with earlier Mesolithic groups. Indigenous communities in fact tended to select specific aspect of the new technology and integrate them with their existing ways of life (Cunliffe 2001). This lead to the presence of two well defined farming groups in the peninsula: a North Italia-Tyrrhenian and a South Italian-Adriatic one (Cunliffe 2001). Later in time, several populations came into contacts with Italian groups, including among the others Greeks, Phoenicians, Arabs, Normans, and Spanish.

These groups settled preferentially on the islands and coastal territories. During the Paleolithic, Italy also received hunter groups coming from Central European areas that the icecap expansion of the LGM was pushing southward. Neanderthal presence is testified from

the numerous skulls and skeletal remains, especially in southern part of Italy. The *Homo sapiens sapiens* presence, going back to the Upper Paleolithic, is testified by the great number of archeological record recovered all around the Italian territory. During the Copper, the Bronze and the Iron ages, few migrations and exchanges occurred between the Mediterranean basin and the Near East. Exchange of metals would determine the transformation of the first social organizations in ancient civilizations. The privilege location of Sardinia, Sicily and Tuscany, coupled with the presence on their territory of important metal resources were the regions firstly inhabited by human groups. Different cultures, recognized on the base of different archeological findings, settlements and burial traditions, provide a precise picture of which probably was the geography of Pre-Historic Italy in the period between the Mesolithic and the Iron Age.

Before the Roman conquest, ancient Italy was characterized just for the presence of Indo-European populations (Pallottino 1981), situated in the Peninsula from the II millennium B.C., above all in the period between the Iron Age and the "Romanization". All these populations are generally known as *italics* (Fig 1.10). The use of this term was parallel to the progressive extension of the geographic boundaries of Italy; at the beginning Italy referred basically to the Calabria region till reaching in successive stages the modern territory known today as Italy. The record of all the populations that inhabited the Italian territory during (pre)-history is naturally incomplete; many of them were of uncertain location and/or ambiguous denomination. Among the most relevant ones are the Veneti in North Oriental; the Latins, Volsci and Aequi in modern Latium (between the Apennine and the West coast); the Sabines, the Umbrians, Marsi, Paeligni, Marrucini and Piceni inhabiting on the other side of the Apennine (above the territories of Umbria and Marche). The Samnites, considered to be an amalgamation of different tribal entities, inhabited in the South Apennine (between modern Abruzzo and Campania). The Osci lived in the inner zones of the actual Campania; while the Yapyges inhabited the area of modern Apulia. Finally, the Lucani and the Brutii occupied the areas of in Basilicata and Calabria. Greek colonies and Phoenicians (e.g. Elymians, Sicani, Siculi) cohabited in Sicily.

Before the rising of the Roman Empire, two non Indo-European populations inhabited in Italy, namely, the Ligures, in the North-West area (between the rivers Arno and Rhone; in a wider area of present Liguria), and the Etruscans with settlements located in

areas far from the Etruria (Tuscany and High Latium) such as Po Plain and the coast of Campania.

For all the VI century B.C., Etruscans represented in Italy the only community with an advanced organization. The origin of the Etruscans, one of the most ancient and enigmatic non-Indo-European civilizations, is being the target of a controversial debate. A recent study identified among modern Tuscans a rather high prevalence of Near Eastern mtDNA haplogroups and an exclusive haplotype sharing between them and Near Eastern populations. The finding has been interpreted as evidence in support of the classical theory that Etruscans may have come from the East through the Mediterranean Sea (Herodotus, Historiae, Vol I, p 94), which currently find little support by archaeologists and historians. In favor of the Eastern Mediterranean origin of the Etruscan civilization, the finding that the extent of mtDNA variation observed in Tuscan cattle breeds is similar to that observed in the Near East and much higher than that observed in the rest of Italy and Europe. The two facts could be compliant with other hypotheses. Thus, studies on fossil DNA in Italy have identified ancient pre-Neolithic bovine – aurochs – whose types are closer to modern bovine than West European aurochs: this contradicts the bovine migration theory and suggests either in loco domestication or population continuity across Italy–Balkans–Anatolia during the Palaeolithic (Achilli et al. 2008). Furthermore, currently available analysis of archaeological Etruscan remains seems to indicate genetic continuity with Tuscans, with closer, but not specific, affinity with Anatolia (Bandelt 2004; Malyarchuck and Rogozin 2004; Vernesi et al. 2004).

Finally, Sardinia, in fact, saw the flourishing of none Indo-European nuragic civilization and, then, the Phoenician colonization.

From Roman period to modern times, Italy has experienced continuous transformations in its demographic composition. The most important italic languages in Italy were of Indo-European origin, beyond Latin, the Umbrian (spoken in the region corresponding to the upper part of the Tiber), and the Oscan, diffused in the Samnitic area, that gave origin to numerous dialects.

After the collapse of the Roman Empire in Europe, the Arab dominance across the Mediterranean was one of the most impressive historical events that occurred in this region. Arabs appeared on the southern shores of the Mediterranean in the early seventh century and quickly conquered North Africa. They spread their language and religion to the native

Northwest African Berber populations, which represented the bulk of the Muslim army that later conquered southern Europe (Davies 1988; Hitti 1990). Referred to either as Moors (in Iberia) or Saracens (in South Italy and Sicily), their arrival in Europe dates to 711 A.D., rapidly subduing most of Iberia and Sicily (831 A.D.). Among European kingdoms their presence was seen as a constant danger, and only by the fifteenth century was the Iberian reconquest completed (Norman 1975). In the thirteenth century Frederick II destroyed Arab rule in Sicily and between 1221 and 1226 he moved all the Arabs of Sicily to the city of Lucera, north of Apulia (Norman 1975). Lucera was later destroyed by Charles II (1301) but an Arab community was recorded in Apulia in 1336 (Norman 1975). Guerrilla warfare was still conducted by Arabs in Sicily even after Frederick II's actions (Capelli et al. 2009; Norman 1975).

Fig 1.10- Pre Roman Italy (arrows indicate Phoenician, Greek and Celtic direction of colonization).

1.7 Ancient DNA

The first aDNA studies used bacterial cloning to amplify small sequences retrieved from skins of animal and human mummies, and revealed the inefficient reaction kinetics of this technique (Higuchi et al. 1984; Pääbo 1985, 1989). These studies demonstrated that the genetic material surviving in ancient specimens was often principally microbial or fungal in origin, and that endogenous DNA was generally limited to very low concentrations of short, damaged fragments of multi-copy loci such as mtDNA. The enormous amplifying power of PCR also created an increased sensitivity to contamination from modern DNA, and simultaneously, a major potential source of such contamination through the extraordinary concentrations of previously amplified PCR products. As a consequence, false positives resulting from intra-laboratory contamination remain a major problem in aDNA research. A series of large scale studies have begun to reveal the true potential of aDNA to record the methods and processes of evolution, providing a unique way to test models and assumptions commonly used to reconstruct patterns of evolution, population genetics and palaeoecological change (Willerslev and Cooper 2005).

aDNA holds tremendous potential for the study of ancient animal and plant populations. Studies on brown bears, penguins, cave bears, horses, dogs and bison (Barnes et al. 2002; Hofreiter et al. 2002; Hofreiter et al. 2004; Leonard et al. 2000; Leonard et al. 2002; Loreille et al. 2001; Ritchie et al. 2004; Vilà et al. 2001) have shown that aDNA can reveal population movements and local extinctions back into the Late Pleistocene. Such studies have considerable power to examine the effects of climate change (e.g. around the LGM) and to test theories and develop methods used in population genetics and palaeobiology. For example, aDNA studies of Beringian brown bears revealed surprisingly large amounts of haplogroup extinction and replacement during the Late Pleistocene and Holocene, and very little interchange of females between populations (Barnes et al. 2002).

The combination of ancient sequences and coalescent methods has considerable power to reconstruct detailed demographic histories, test models of population genetics and reveal much novel information about microevolutionary processes. These methods can even recover demographic data for taxa that have been through recent population bottlenecks, which would normally remove genetic signals (Shapiro et al. 2004). Such methods also provide an opportunity to directly estimate evolutionary rates of nucleotide substitution and

directly date phylogenetic events without using an external paleontological calibration for a molecular clock (Lambert et al. 2001; Ritchie et al. 2004; Shapiro et al. 2004). The results show that palaeontologically calibrated rate estimates are often significantly slower than those calibrated from aDNA population studies and may reflect differences in sequence substitution processes operating at short and long timeframes (Willerslev and Cooper 2005).

Objectives

2. **Objectives**

The genetic variability of humans is the common ground upon which the two disciplines of population genetics and forensic genetics are based. The former studies the genetic variability associated to groups of individuals, i.e. populations; the latter uses this variability, both in criminal cases and in paternity tests, for the characteriztion of individuals, through a genetic ´fingerprinting`.

Population genetic studies the distribution of genetic diversity among populations. Since the genomic variation of living individuals is the result of the human evolutionary process, such a study can provide information on past demographic events.
Thus, the main aims of this thesis are:

1. describing the distribution of genetic diversity beard by modern human populations and of its apportionment among subpopulations, i.e. population structure, focusing on populations from Sub-Saharan Africa and Europe, with a special focus on Cameroon, Western and Central Africa, and Italy, respectively. We made use of information compiled from previous studies and as well as historical, linguistic and geographic data.
2. inferring the pre-historical and historical events that determined the observed modern diversity and structure. The inference process, based on population genetics model, is strongly motivated by anthropological interest in the history of our species, its origins, movement and demographic development.

In pursuing these global aims, this thesis is divided in three main groups: two geographical groups, including Africa and Europe, and one group for ancient DNA. The intermediary objectives are so defined:

i) **Sub-Saharan Africa**:

1) to evaluate the genetic variation of autosomal and Y-chromosomal microsatellites in a large set of Cameroon samples, also taking geographic and cultural factors into consideration.

2) to genetically characterize the Y chromosome variation in order to reconstruct the demographic events and identify specific lineages associated with the spread of languages, agriculture, and pastoralism in sub-Saharan Africa.

ii) Europe:

1) to perform a phylogeographic analysis of mtDNA variation at the highest level of molecular resolution, trying to explain how the refuge area located on the Italian Peninsula contributed to the human recolonization of the continent at the beginning of the Holocene.
2) to investigate the role of the Italian Peninsula as part of a more global process of the peopling of Europe, considering the demographic consequences of the agriculture revolution in the area by genotyping Y chromosome markers for a large number of samples.
3) to evaluate the origin of Y chromosome haplogroup R1b1b2-M269, investigating the frequency patterns and diversity in a large chromosomes collection yet assembled.
4) to analyse the genetic patterns of Italy from a global perspective, using 12 different populations along the Italian Peninsula, two of them being linguistic isolates, and analyzed for the mtDNA control region and selected coding region SNPs, a panel of Y-chromosome SNPs and STRs and, in addition, for autosomal AIMs.
5) to investigate the Etruscan origins (they are among one of the most enigmatic non-Indo-European civilizations) through the analysis of modern Tuscans and using mtDNA SNPs and complete genome sequencing.
6) to determine the higher recent Northwest African male legacy contribution in Iberia and Sicily.
7) to contribute to enrich the Y-chromosome databases regarding high-resolution Y-chromosome data sets.
8) to evaluate the genealogical correlation between the Y-chromosomes carrying the DYS19 microsatellite duplication.
9) to contribute on the autosomal allele frequencies of different Italian populations.

iii) Ancient DNA:

1) to study the wooly mammoth (*Mammuthus primigenius*) using mtDNA markers in order to resolve is phylogenetic affiliation within Elephantidae.

Results

3. **Results**

Due to the wide-range of topics presented in the present thesis document, we have decided to group the different articles following a more logical structure that do not necessarily follows a chronological order. This is the same structure we have tried to follow along the rest of the thesis document, including Discussion and Conclusions sections.

3.1 Sub-Saharan Africa

Article 1. A multi-perspective view of genetic variation in Cameroon. *Am J Phys Anthropol*

Article 2. Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Mol Biol Evol*

AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 140:454–464 (2009)

A Multi-Perspective View of Genetic Variation in Cameroon

V. Coia,[1,2] F. Brisighelli,[3,4] F. Donati,[1] V. Pascali,[3] I. Boschi,[3] D. Luiselli,[5] C. Battaggia,[1] C. Batini,[1,6] L. Taglioli,[7] F. Cruciani,[8] G. Paoli,[7] C. Capelli,[9] G. Spedini,[1,10] and G. Destro-Bisol[1,10]*

[1]*Dipartimento di Biologia Animale e dell'Uomo, Sapienza Università di Roma, Roma 00185, Italia*
[2]*Dipartimento di Filosofia, Storia e Beni Culturali, Università di Trento, Trento 38100, Italia*
[3]*Istituto di Medicina Legale, Università Cattolica di Roma, Roma 00168, Italia*
[4]*Unidade de Xenética, Instituto de Medicina Legal, Universidade de Santiago de Compostela, Santiago 15782, Spain*
[5]*Dipartimento di Biologia Evoluzionistica Sperimentale, Università di Bologna, Bologna 40126, Italia*
[6]*Department de Ciències Experimentals i de la Salut, University "Pompeu Fabra, Barcelona 08002, Spain*
[7]*Dipartimento di Biologia, Università di Pisa, Pisa 56126, Italia*
[8]*Dipartimento di Genetica e Biologia Molecolare, Sapienza Università di Roma, Roma 00185, Italia*
[9]*Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK*
[10]*Istituto Italiano di Antropologia, Roma 00185, Italia*

KEY WORDS microsatellites; autosomes; Y-chromosome; mtDNA; language; geography

ABSTRACT In this study, we report the genetic variation of autosomal and Y-chromosomal microsatellites in a large Cameroon population dataset (a total of 11 populations) and jointly analyze novel and previous genetic data (mitochondrial DNA and protein coding loci) taking geographic and cultural factors into consideration. The complex pattern of genetic variation of Cameroon can in part be described by contrasting two geographic areas (corresponding to the northern and southern part of the country), which differ substantially in environmental, biological, and cultural aspects. Northern Cameroon populations show a greater within- and among-group diversity, a finding that reflects the complex migratory patterns and the linguistic heterogeneity of this area. A striking reduction of Y-chromosomal genetic diversity was observed in some populations of the northern part of the country (Podokwo and Uldeme), a result that seems to be related to their demographic history rather than to sampling issues. By exploring patterns of genetic, geographic, and linguistic variation, we detect a preferential correlation between genetics and geography for mtDNA. This finding could reflect a female matrimonial mobility that is less constrained by linguistic factors than in males. Finally, we apply the island model to mitochondrial and Y-chromosomal data and obtain a female-to-male migration Nv ratio that was more than double in the northern part of the country. The combined effect of the propensity to interpopulational admixture of females, favored by cultural contacts, and of genetic drift acting on Y-chromosomal diversity could account for the peculiar genetic pattern observed in northern Cameroon. Am J Phys Anthropol 140:454–464, 2009.

The area occupied by the present-day Republic of Cameroon is of particular importance for bio-anthropological studies, since it may be regarded as a sort of transect that comprises an important part of the vast biological and cultural diversity of sub-Saharan Africa (see Campbell and Tishkoff, 2008 for a review). In fact, it is inhabited by a large wealth of populations, which differ substantially in subsistence strategies, language, social structure, and religion. A primary distinction can be made using a geographic criterion. The populations living in the northern part of the country (provinces *Extrême Nord* and *Nord*), where the Savannah and Sahel habitat predominates, are often referred to as "Sudanese" populations. Their typical subsistence economy (the so called "Sudanic complex") is mainly based on the cultivation of cereals such as kaffir corn and millet (*Sorghum* and *Panicum miliaceum* L.) and the breeding of cattle and sheep (Harris, 1976; Ehret, 1984). The Sudanese are linguistically heterogeneous, speaking languages belonging to the Afro-Asiatic, Nilo-Saharan and Niger-Kordofan phyla (Greenberg, 1980). Some of these populations, the so-called *Montagnards* (Podokwo, Mada, and Uldeme), are thought to be in continuity with the groups that created the oldest nucleus of settlements in northern Cameroon, the "Sao civilization" (Lebeuf, 1981; McIntosh and McIntosh, 1983). The northern region was first populated around 8,000 BP by groups belonging to the original Afro-Asiatic or Hamito-Semitic linguistic stock from the Near East (David, 1981). Subsequently, another migration of people from the Sahelian area to the South reached the region of the Adamawa plateau around 4,000 BP, a likely consequence of the desertifica-

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Sapienza Università di Roma (ricerche di Ateneo); Istituto Italiano di Antropologia, Roma, Italia.

*Correspondence to: Giovanni Destro-Bisol, Department of Animal and Human Biology, University "La Sapienza," P.le A. Moro 5, 00185 Roma, Italia. E-mail: destrobisol@uniroma1.it

Received 18 October 2008; accepted 6 March 2009

DOI 10.1002/ajpa.21088
Published online 7 May 2009 in Wiley InterScience (www.interscience.wiley.com).

tion of the Sahara (5,000–7,000 BP) (David, 1981). Finally, the area was recently populated (18th century) by the Fulbe from Nigeria (Mohammadou, 1973).

The climate of the southern land (provinces *Ouest*, *Littoral*, and *Centre*) is equatorial and most of the territory is occupied by tropical and equatorial forest. Some of the populations settled in this area are thought to be descendants of the proto-Bantu nucleus that originated on the Nigerian-Cameroon plateau around 4,000 BP (Bakaka and Bassa), whereas others (Bamileke and Ewondo) settled more recently (18th century) in the area and adopted a Bantu language (Ehret, 1984; Spedini et al., 1999; see below). Their traditional agricultural techniques, the so-called "vegecultural complex" (Harris, 1976), use tubers including yam and manioc, while the diet is complemented by the breeding of small animals and poultry. In contrast to north Cameroon, populations from south are linguistically homogeneous. In fact, they speak languages of the sub-branch Benuè-Congo of the Niger-Kordofan phylum (Greenberg, 1980), generally referred to as Bantu.

Given this complex background, Cameroon provides a unique opportunity to study how biological, geographic, and cultural factors interact in determining variation within and among human populations.

The genetic variation of Cameroon populations has been analyzed at protein coding loci (Spedini et al., 1999), mitochondrial DNA (Cerný et al., 2004, 2007; Destro-Bisol et al., 2004a,b; Coia et al., 2005), and Y-chromosome (Cruciani et al., 2002; Caglià et al., 2003; Coia et al., 2004; Wood et al., 2005). However, there is limited data regarding microsatellite polymorphisms, which are particularly useful for population genetic studies due to their high level of variation.

This study analyzes the genetic variation of autosomal and Y-chromosomal microsatellites (16 and 6 loci, respectively). We also compare the new results with available genetic data (concerning mtDNA and protein coding loci), with the aiming to achieve a more complete view of the genetic structure of Cameroon populations.

MATERIALS AND METHODS

Populations

The dataset includes a total of eleven populations (see Fig. 1). Following Greenberg's linguistic classification (Greenberg 1963, 1980; Barreteau et al., 1984), northern Cameroon is represented by four Afro-Asiatic (Chadic sub-branch, Mada, Mandara, Podokwo, and Uldeme) and two Niger Kordofanian speaking populations (Adamawa sub-branch, Fali, Tupuri; West Atlantic sub-branch, Fulbe). In southern Cameroon, four Bantu-speaking populations were studied (Niger Congo, Benuè-Congo sub-branch, Bakaka, Bamileke, Bassa, and Ewondo). A subset of data for autosomal microsatellites studied here and complete Y-chromosomal typings for Bamileke and Ewondo have been previously published (Destro-Bisol et al., 2000; Caglià et al., 2003).

Laboratory analyses

We analyzed variation at 16 autosomal microsatellites in a total of 454 apparently healthy and unrelated individuals (see Fig. 1). After extraction from whole blood samples by the salting-out method (Miller et al., 1988), the DNA samples were typed for thirteen loci of the Combined DNA Index System (CODIS; Butler, 2006)

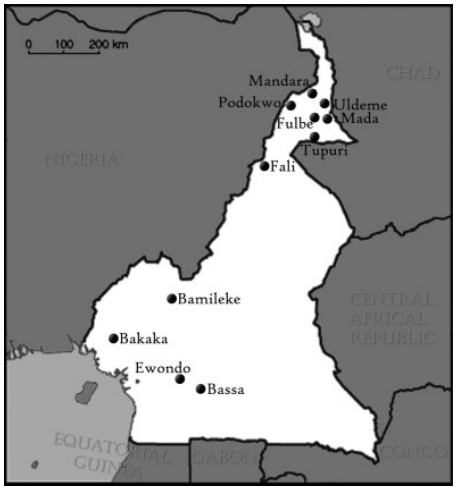


Fig. 1. Geographic location of the populations analyzed in the study.

(D18S51, D21S11, TH01, D3S1358, FGA, TPOX, D8S1179, vWA, CSF1PO, D16S539, D7S820, D13S317, D5S818) plus HumCD4, HumFES, and HumF13A1. The genotyping of the CODIS loci plus the Amelogenin locus for sex typing was carried out in the laboratory of Forensic Haematology of the Catholic University of Rome, using AmpF/STR Profiler Plus, AmpF/STR COfiler (Applied Biosystem) and PowerPlex-16 (Promega) PCR amplification kits according to product instructions.

Fragment sizes were detected by the ABI PRISM 310 genetic Analyzer (Applied Biosystem) using Genescan 500 ROX as an internal-size standard and sequenced allele ladders. The genotyping of HumCD4, HumFES, and HumF13A1 loci was performed in the laboratory of Molecular Anthropology of the University of Rome "La Sapienza." PCR conditions were as previously described (Destro-Bisol et al., 2000). Allele nomenclature follows the recommendations of the European DNA profiling group (EDNAP; Gill et al., 1997).

Eight out of the 11 populations studied for autosomal variation were also analyzed for Y-chromosome microsatellite variation (a total of 281 samples). Mada were excluded from Y-chromosome typing due to the insufficient number of males, while Bamileke and Ewondo have been analyzed previously (Destro-Bisol et al., 2000; Caglià et al., 2003). The six Y-chromosome microsatellites (DYS19, DYS389-I, DYS390, DYS391, DYS392, and DYS393) were typed according to the conditions described by Kayser et al. (1997). The amplified products were separated on polyacrylamide denaturing gels (7 M urea, 7% T) using a semi-automated DNA sequencer (A.L.F. Express, Pharmacia Biotech, Uppsala, Sweden) or by capillary electrophoresis using the ABI PRISM 310 sequencer (Applied Biosystem) (Caglià et al., 2003). Allelic and internal standards were used for microsatellite typing (Moscetti et al., 1995). Allele nomenclature was standardized according to Kayser et al. (2001).

TABLE 1. Intra-population diversity parameters in the 11 populations analyzed for autosomal microsatellite variation

Population	2n	AGD	SE	t.n.a	n.p.a.	DHW	HDT	HET
North								
Fali	66	0.767	0.016	123	1	D5S818 (P = 0.004)	D5S818 (P = 0.007); D18S51 (P = 0.005)	F13A1 (***P* = 0.002**)
Fulbe	78	0.786	0.013	127	2	F13A1 (***P* = 0.000**)		
Mada	80	0.770	0.017	133	3	CD4 (P = 0.010); D16S539 (P = 0.006)	D16S539 (P = 0.011)	
Mandara	50	0.773	0.019	124	2		D21S11 (P = 0.009)	
Podokwo	82	0.760	0.020	125	4	FES (P = 0.020)	D21S11 (P = 0.038); FES (P = 0.030)	
Tupuri	50	0.752	0.018	117	2	FGA (P = 0.039); D13S317 (P = 0.048)	D5S818 (P = 0.004); FGA (P = 0.029)	
Uldeme	92	0.763	0.018	130	3	FGA (P = 0.004); D13S317 (P = 0.031)	CD4 (P = 0.043); D13S317 (P = 0.048)	
South								
Bamileke	60	0.777	0.018	123	3			
Bakaka	116	0.770	0.016	130	2		vWA (P = 0.040); D8S1179 (P = 0.037)	
Bassa	116	0.773	0.016	127	1	D18S51 (P = 0.037)	D18S51 (P = 0.010)	
Ewondo	118	0.793	0.014	142	4	TPOX (P = 0.018); D21S11 (P = 0.019)	D21S11 (P = 0.027)	

Significant P values, after Bonferroni correction, are reported in boldface ($P < 0.0031$). 2n, number of chromosomes; AGD, average gene diversity and its standard error (ES); t.n.a, total number of alleles; n.p.a., number of private alleles; DHW, departures from Hardy Weinberg equilibrium; HDT, heterozygosity deficiency test; HET, heterozygosity excess test.

Statistical analyses

Allele frequencies were obtained directly by gene-counting. The intra-population diversity parameters for autosomal loci were obtained using the GENEPOP software, ver. 4 (Rousset, 2008). Unbiased heterozygosity and standard error for single autosomal microsatellite loci were estimated using equations 8.4 and 8.12 described in Nei (1987). Haplotype diversity (HD), mean number of pairwise differences (MNPDs) and number of different haplotypes (h) for unilinearly transmitted loci were calculated using the Arlequin software, ver. 3.1 (Excoffier et al., 2005).

Departures from Hardy-Weinberg Equilibrium (HWE) were evaluated using a two-tailed (probability test) and one-tailed (heterozygote excess and deficiency) exact tests with a Markov chain simulation method as implemented in the GENEPOP software package, ver. 4 (Rousset, 2008).

Genetic differentiation among populations was estimated through F_{st} genetic distances (Reynolds, 1983). The genetic distances were visualized in a Multi-Dimensional Scaling (MDS) plot (Kruskal, 1964) obtained by using the STATISTICA software (StatSoft, Inc., 1997). The significance of the stress value was evaluated according to Sturrock and Rocha (2000). To analyze the relationships between genetic distances (F_{st}) and geographic and linguistic distance matrices, we computed Spearman's rho correlation and partial correlation coefficients using Tanagra software (Rakotomalala, 2005). This method was preferred to Pearson's product moment correlation r since it assumes only a monotonic relationship (as opposed to a linear one) between the variables (Pollard, 1977; Relethford, 1985). Geographic distances were calculated in kilometers by measuring the road distances between the main villages. We assigned the linguistic distances as follows: $d = 0$ for populations belonging to the same sub-branch; $d = 1$ for the same branch; $d = 2$ for the same phylum and $d = 3$ for a different phylum. Probability values were assessed using the Mantel's test (Relethford, 1990, http://konig.la.utk.edu/relethsoft.html). As suggested by Jackson and Somers (1989) the number of permutations was fixed at 10,000 to minimize the fluctuation of probability values.

The Analysis of Molecular variance (AMOVA) (Excoffier et al., 1992) was performed using the software Arlequin, ver 3.1 (Excoffier et al., 2005; see also Castrì et al., 2008).

In order to detect undeclared relationships, we used a likelihood approach as implemented in the KINGROUP program (Konolalov et al., 2004). The program calculates and compares the likelihood of the primary hypothesis (presence of genetic relatedness) and the null hypothesis, H_0 (absence of genetic relatedness), providing a P value for the acceptance of the primary hypothesis. We tested four possible levels of relationship (parents-offspring, full siblings, half sibling and cousins) using the 16 autosomal markers available.

RESULTS

Variation at microsatellite loci

Autosomal microsatellites. Allele frequencies and single-locus gene diversity values are available in the Supporting Information Material (Supporting Information Tables S1 and S2). Intra-population diversity parameters, including the total number of different alleles and private alleles detected at all loci in each population are reported in Table 1.

Tupuri shows the smallest number of different alleles (117), and the Ewondo the largest (142). Private alleles are comprised between one (Fali) and four (Ewondo and Podokwo), with frequencies ranging from 0.8% (allele 16 at F13A1, Ewondo) to 4.9% (allele 29.2 at D21S11, Podokwo).

Departures from the Hardy-Weinberg equilibrium (DHW) were detected at 10 loci (CD4, D5S818, D13S317, D16S539, D18S51, D21S11, FES, FGA, F13A1, and TPOX) in 8 out of the 11 populations analyzed (Bassa, Ewondo, Fali, Fulbe, Podokwo, Tupuri, Uldeme, and Mada) using the Fisher exact test (Table 1). The single-locus heterozygosity deficiency test (HDT) detected statistically significant departures ($\alpha \leq 0.05$) at nine loci

(CD4, D8S1179, D13S317, D16S539, D18S51, D21S11, FES, FGA, and vWA) occurring in nine populations. One case of excess of heterozygosity (HET) was observed (locus F13A1, Fali population). However, the importance of these departures should not be overvalued since only two of them remained significant after the application of the Bonferroni correction (DHW, F13A1 Fulbe; HET, F13A1 Fali) (Table 1).

Y-chromosomal microsatellites. Allele frequencies are available in the Supporting Information Material (Supporting Information Table S3). Eighty-three different haplotypes were detected, 63 of which were found in only one population (76% of the total; Table 2). The number of different haplotypes ranges from 4 (Podokwo) to 22 (Fulbe and Mandara). The minimum number of private haplotypes was found in Podokwo (2), while the maximum was found in Fulbe and Mandara (13).

Our results were compared with the YHRD STR database (http://www.yhrd.org) which includes data on 499 world-wide populations, 28 of which from Africa (a total of 3,143 haplotypes including 1,126 from sub-Saharan Africa). Twenty private haplotypes for our dataset (32% of total private haplotypes) were not detected in the other African populations, but were observed in Europeans and Asians. Most of them (75%) belong to the northern Cameroon populations, where they are present at low frequencies with the exception of haplotypes H36 in Fali (frequency of 28%) and H50 in Uldeme (frequency of 46%). Moreover, 7 out of the 63 haplotypes defined as private using our dataset, are virtually restricted to Cameroon (H6, H17, H21, H28, H48, H53, and H72). They were all found in the North, the only exception being one haplotype found among Bassa (H21). Finally, the most common haplotype found in each population (Supporting Information Table S4) was also checked for matches and no identical types were found in Africa for Podokwo (H63) and Uldeme (H50). On the whole, Y-chromosome gave a strong signal of genetic distinctiveness of northern Cameroon populations from the other African populations, which was obtained despite a relatively low number of Y-STR loci.

The haplotype diversity values of Podokwo (0.276 ± 0.109), Uldeme (0.752 ± 0.070) and Tupuri (0.677 ± 0.106) are remarkably low, with the value observed in Podokwo accounting for one third of the estimates of the least diverse populations from South Cameroon (Ewondo; 0.829 ± 0.044). Interestingly, a comparable reduction of Y-chromosomal diversity at loci DYS19, DYS389-I, DYS390, DYS391, DYS392, and DYS393 has so far been observed only in Central Asian shepherds (Chaix et al., 2007). The result obtained in Podokwo and Uldeme is paralleled by a reduced haplogroup variability in both populations, since the R1b1* (xR1b1b2) haplogroup is highly prevalent in these populations (frequency of 85% in Uldeme and 95% in Podokwo; Cruciani et al., 2002; Wood et al., 2005). The distribution of Y-chromosome MNPD values shows a comparable pattern, with reduced values for the Podokwo (0.575 ± 0.480) and Uldeme (1.761 ± 1.050) compared to other populations. On the other hand, a less marked prevalence of the R1b1* (xR1b1b2) haplogroup (66%) and no substantial reduction in the MNPD value (2.311 ± 1.307) was observed in Tupuri. In order to test the hypothesis that the observed extremely reduced Y-chromosome variability in Podokwo and Uldeme could be due to undeclared family relationships among donors, we used a likelihood approach as implemented in the KINGROUP program (Konolalov et al., 2004). In the parents-offspring relationships, one Podokwo pair and one Uldeme pair were detected, but only the Podokwo pair included two males, probably linked through a father–son relationship. Y-chromosome analysis confirmed an identical 6-STR loci haplotype. For the other three hypotheses (full siblings, half sibling, and cousins), only one Podokwo pair was found to be significant. However, it was composed of two females and had no implication for analysis of Y-chromosome variation. The HD and the MNPD were recalculated in Podokwo excluding one individual from the pair with a possible father–son relationship. The values increased only slightly, confirming the very low levels of intra-population diversity (HD 0.2862 ± 0.1121 and MNPD 0.5969 ± 0.4923).

Comparing new and old data

We compared the data obtained in the course of this study with previously published results regarding mtDNA (hypervariable region-1, np 16024-16390; Destro-Bisol et al., 2004; Coia et al., 2005). Furthermore, specifically for correlation and inter-population diversity analyses we considered the data from 10 polymorphisms of protein coding loci (PCL) (6-PGD, ACP, CAII, ESD, GLO, GPX1, PGM1, A1-AT, GC and TF published by Spedini et al. (1999), which represents to date the reference study of genetic variation in Cameroon on a regional scale.

The populations from southern Cameroon (all Bantu speakers) are more homogeneous for all intra-population diversity parameters (Y-chromosome HD and MNPDs, mtDNA HD), the only exception being mtDNA MNPDs (Tables 3 and 4). However, the results of Y-chromosomal microsatellites are the most noticeable. In fact, the variance of HD values is significantly higher in northern than southern populations (0.0573 vs 0.000349, $P < 0.001$). The distribution of Y-chromosome MNPD values shows a comparable pattern, with a greater fluctuation in Cameroon (variance 0.9850 vs 0.0984, $P < 0.05$). In sharp contrast with Y-chromosome, no mtDNA reduction of HD or MNPD was detected in the Podokwo and Uldeme, whose values are among the highest observed in Cameroon and in sub-Saharan Africa (see Salas et al., 2002). An explanation of this difference between the two unilinearly transmitted loci could lie in differences between female and male migration rates and/or effective size (see Discussion).

The analysis of inter-population diversity also points to a substantial differentiation between the two geographic groups. In fact, the F_{st} for northern Cameroon largely exceeds those obtained for the southern part for all the genetic systems analyzed. The difference is greater for Y-chromosomal microsatellites (0.361 vs 0.156) than for autosomal microsatellites (0.013 vs 0.006), PCL (0.010 vs 0.007) or mtDNA (0.021 vs 0.016). Even when the outliers (Podokwo and Uldeme) are excluded from the analysis (see Fig. 2), the greater variation of northern Cameroon is confirmed, with the highest value found again for the Y-chromosome (0.235 vs 0.155; autosomal microsatellites, 0.010 vs 0.006; PCL 0.008 vs 0.007; mtDNA 0.025 vs 0.016).

The plots of the F_{st} genetic distances among populations are reported in Figure 2. The northern and southern populations are found in different parts of the plots, with the former being more separated from each other

TABLE 2. Y-chromosome microsatellite haplotypes in the 10 populations analyzed

Haplotype	DYS19	DYS389-I	DYS390	DYS391	DYS392	DYS393	FAL	FUL	MAN	POD	TUP	ULD	BAK	BAS	*n*	*F*
H1	13	12	22	9	11	13	0	0	1	0	0	0	0	0	1	0.003
H2	13	12	24	10	11	12	0	1	0	0	0	0	0	0	1	0.003
H3	13	13	24	10	11	12	0	0	0	1	0	0	0	0	1	0.003
H4	14	12	24	10	14	14	0	1	0	0	0	0	0	0	1	0.003
H5	14	12	25	11	11	13	0	0	4	0	0	0	0	0	4	0.014
H6	14	13	20	11	11	13	0	0	0	0	1	0	0	0	1	0.003
H7	14	13	21	10	11	13	0	1	0	0	0	0	0	0	1	0.003
H8	14	13	21	11	11	13	0	2	0	0	0	0	0	0	2	0.007
H9	14	13	24	11	13	13	0	1	0	0	0	0	0	0	1	0.003
H10	14	13	25	11	11	13	0	0	1	0	0	0	0	0	1	0.003
H11	14	14	23	11	13	13	0	0	0	0	1	0	0	0	1	0.003
H12	14	14	24	11	13	13	0	0	1	0	0	0	0	0	1	0.003
H13	15	12	21	9	11	13	0	1	0	0	0	0	0	0	1	0.003
H14	15	12	21	9	12	13	0	1	0	0	0	0	0	0	1	0.003
H15	15	12	21	10	11	14	0	0	1	0	0	0	2	0	3	0.010
H16	15	12	22	9	12	13	0	1	0	0	0	0	0	0	1	0.003
H17	15	12	22	11	15	13	0	0	1	0	0	0	0	0	1	0.003
H18	15	13	20	10	11	14	0	1	0	0	0	0	0	0	1	0.003
H19	15	13	21	9	11	13	0	0	0	0	0	0	0	1	1	0.003
H20	15	13	21	9	11	14	0	0	0	0	0	0	0	16	16	0.056
H21	15	13	21	9	11	15	0	0	0	0	0	0	0	1	1	0.003
H22	15	13	21	10	11	12	1	0	0	0	0	0	0	0	1	0.003
H23	15	13	21	10	11	13	4	2	0	0	1	0	4	2	13	0.046
H24	15	13	21	10	11	14	1	2	1	0	0	0	4	5	13	0.046
H25	15	13	21	10	11	15	0	0	0	0	1	0	0	1	2	0.007
H26	15	13	21	11	11	13	0	1	1	0	0	0	0	1	3	0.010
H27	15	13	21	11	11	14	0	0	0	0	0	0	1	2	3	0.010
H28	15	13	21	11	13	14	0	1	0	0	0	0	0	0	1	0.003
H29	15	13	22	10	11	13	0	1	0	0	0	0	0	0	1	0.003
H30	15	13	23	10	11	13	1	0	0	0	0	0	0	0	1	0.003
H31	15	13	23	10	13	13	0	0	0	0	1	0	0	0	1	0.003
H32	15	13	23	10	13	14	0	0	0	0	1	0	0	0	1	0.003
H33	15	13	23	11	13	13	0	0	1	2	15	0	0	0	18	0.064
H34	15	13	24	10	11	13	2	0	0	0	0	5	0	0	7	0.024
H35	15	13	24	10	13	13	2	1	1	0	0	0	0	0	4	0.014
H36	15	13	24	10	14	13	10	0	0	0	0	0	0	0	10	0.03
H37	15	13	24	11	13	13	0	1	1	0	0	0	0	0	2	0.007
H38	15	13	24	11	14	13	0	0	0	0	0	4	0	0	4	0.014
H39	15	13	24	11	14	14	0	0	0	0	0	1	0	0	1	0.003
H40	15	14	21	9	11	14	0	0	0	0	0	0	0	1	1	0.003
H41	15	14	21	10	11	12	1	0	0	0	1	0	0	0	2	0.007
H42	15	14	21	10	11	13	1	1	0	0	0	0	0	0	2	0.007
H43	15	14	21	10	11	14	0	1	0	0	0	0	0	0	1	0.003
H44	15	14	23	10	13	13	0	0	2	0	0	0	0	0	2	0.007
H45	15	14	23	10	15	13	0	0	1	0	0	0	0	0	1	0.003
H46	15	14	23	11	13	13	0	0	3	0	0	0	0	0	3	0.010
H47	15	14	23	11	13	14	0	0	1	0	0	0	0	0	1	0.003
H48	15	14	23	12	14	13	0	0	0	0	0	1	0	0	1	0.003
H49	15	14	24	11	13	13	0	0	2	0	0	1	0	0	3	0.010
H50	15	14	24	11	14	13	0	0	0	0	0	14	0	0	14	0.049
H51	15	14	25	11	13	13	0	0	0	0	0	0	0	1	1	0.003
H52	15	15	23	11	13	13	0	0	0	0	0	1	0	0	1	0.003
H53	15	15	24	11	15	13	0	0	0	0	0	2	0	0	2	0.007
H54	16	12	24	10	11	13	1	0	0	0	0	0	0	0	1	0.003
H55	16	13	21	10	11	12	1	0	0	0	0	0	0	0	1	0.003
H56	16	13	21	10	11	13	1	0	0	0	1	0	5	2	9	0.032
H57	16	13	21	10	11	14	0	3	0	0	0	0	2	4	9	0.032
H58	16	13	21	10	11	15	1	0	0	0	0	0	0	1	2	0.007
H59	16	13	21	11	11	13	0	0	0	0	0	0	14	0	14	0.049
H60	16	13	21	11	13	15	0	0	1	0	0	0	0	0	1	0.003
H61	16	13	22	10	11	13	0	0	0	0	1	0	0	0	1	0.003
H62	16	13	23	10	13	13	0	0	0	1	0	0	0	0	1	0.003
H63	16	13	23	11	13	13	0	0	2	23	0	0	0	0	25	0.088
H64	16	13	24	10	11	11	2	0	0	0	0	0	0	0	2	0.007
H65	16	13	24	10	11	13	13	1	0	0	0	1	0	0	15	0.053
H66	16	13	24	11	13	14	0	0	0	0	1	0	0	0	1	0.003
H67	16	14	21	10	11	12	0	0	1	0	0	0	0	0	1	0.003
H68	16	14	25	10	11	13	0	0	0	0	0	0	0	1	1	0.003
H69	16	14	21	10	11	14	0	1	0	0	0	0	0	1	2	0.007

(continued)

TABLE 2. (Continued)

Haplotype	DYS19	DYS389-I	DYS390	DYS391	DYS392	DYS393	FAL	FUL	MAN	POD	TUP	ULD	BAK	BAS	*n*	*F*
H70	16	14	21	10	11	15	0	0	0	0	0	0	0	2	2	0.007
H71	16	14	21	11	11	13	0	0	0	0	0	0	1	0	1	0.003
H72	16	14	21	11	11	14	0	0	1	0	0	0	0	0	1	0.003
H73	16	14	21	11	11	15	0	0	0	0	0	0	0	1	1	0.003
H74	16	14	24	11	13	14	0	0	0	0	0	0	0	1	1	0.003
H75	17	12	22	9	12	13	0	1	0	0	0	0	0	0	1	0.003
H76	17	13	21	10	11	14	0	0	0	0	0	0	8	0	8	0.028
H77	17	13	21	10	11	15	0	0	0	0	0	0	6	0	6	0.021
H78	17	13	23	11	13	13	0	0	1	0	0	0	0	0	1	0.003
H79	17	13	24	10	11	14	1	0	0	0	0	0	0	0	1	0.003
H80	17	14	21	10	11	14	0	0	1	0	1	0	0	0	2	0.007
H81	17	14	21	10	11	15	0	0	0	0	0	0	0	5	5	0.017
H82	17	15	21	10	11	15	0	0	0	0	0	0	1	0	1	0.003
H83	18	13	21	10	11	15	0	0	0	0	0	0	1	0	1	0.003

FAL, Fali; FUL, Fulbe; MAN, Mandara; POD, Podokwo; TUP, Tupuri; ULD, Uldeme; BAK, Bakaka; BAS, Bassa; *n*, number of individuals observed for each haplotype; *F*, frequency for each haplotype.

TABLE 3. Intra-population diversity parameters calculated in the 10 populations analyzed for three genetic systems (autosomal microsatellites STRs, Y-chromosome microsatellites Y-STRs and mitochondrial DNA)

	STRs			Y-STRs				mtDNA			
Population	*N*	AGD	SE	*N*	*h*	HD	SE	*N*	*h*	GD	SE
North											
Fali	33	0.773	0.028	43	16	0.854	0.038	41	26	0.977	0.010
Fulbe	39	0.789	0.023	27	22	0.983	0.015	34	26	0.975	0.016
Mandara	25	0.781	0.034	30	22	0.972	0.017	37	31	0.990	0.009
Podokwo	41	0.765	0.027	27	4	0.276	0.109	39	33	0.991	0.008
Tupuri	25	0.760	0.038	26	12	0.677	0.106	25	23	0.993	0.013
Uldeme	46	0.767	0.024	30	9	0.752	0.070	28	25	0.992	0.012
South											
Bamileke	30	0.783	0.030	50	19	0.876	0.037	48	36	0.988	0.007
Bakaka	58	0.773	0.021	49	12	0.866	0.028	50	36	0.983	0.008
Bassa	58	0.776	0.019	49	19	0.872	0.038	46	39	0.992	0.007
Ewondo	59	0.797	0.018	39	14	0.829	0.044	53	39	0.983	0.008

N, number of samples; HD, haplotype diversity; AGD, average gene diversity; *h*, number of haplotypes.

than the latter. According to the F_{st} estimates, the separation is more evident for Y-chromosome microsatellites, where southern populations are tightly clustered and the northern ones spread in the bi-dimensional space. Taking the statistical significance of genetic distances into account, the genetic systems produced different results. All genetic distances calculated for Y-chromosome and most of those based on autosomal STRs (41 out of 45) and PCL (40 out of 45) are statistically significant ($P < 0.05$). In the case of mtDNA, most of values obtained comparing south and north populations are statistically significant (21 out of 24), while the majority of distances between populations from the same geographic area are insignificant (5 out of 6 for the southern group; 9 out of 15 for the northern group). This indicates a more robust geographic structure of genetic diversity for mtDNA, which is consistent with AMOVA and correlation analyses (see below).

The AMOVA analysis was carried out either on the entire dataset or dividing it into two geographic groups (northern and southern Cameroon) or into two (Afro-Asiatic and Niger-Kordofan) or four [(Afro-Asiatic (Chadic), Niger-Kordofan (Adamawa), Niger-Kordofan (Benué-Congo) and Niger-Kordofan (West-Atlantic)] linguistic groups. With all datasets, both autosomal microsatellites and PCL show a low value of variation among populations (0.8–1.11% and 0.7–1.01% of total variation), with more than 98% of the variation within populations (Table 5). A slightly higher value was obtained for mtDNA (1.80–2.94%). A striking difference was observed for Y-chromosome, which shows the highest level of differentiation among populations (22.0–34.8%). The difference remains substantial even when the diverging populations of Podokwo (28.8%) or Uldeme (33.3%) were excluded from the entire dataset (see Fig. 2B).

The strongest signals of differentiation among groups are provided by the unilinearly transmitted polymorphisms, which produced values from 3 to 100 times higher than other genetic systems (Table 5). Taking both the magnitude of values and their statistical significance into account, geography seems to be the best predictor of among-group variation for both systems (Y-chromosome 16.1%; mtDNA 2.1%). On the other hand, language seems to be more important for determining among-group variation of autosomal STRs and PCL, but the values produced are extremely low (0.4% STRs; 0.6% and 0.5% PCL, see Table 5). Correlation analyses are reported in Table 6. The values between mtDNA and geography or PCL and language are the only which remain appreciable (i.e. with a proportion of variance explained, $r^2 < 20\%$) and statistically significant when partial correlations are used.

TABLE 4. Mean number of pairwise differences (MNPD) for the Y-chromosome microsatellites and mitochondrial DNA in the 10 populations analyzed

	Y-STRs				mtDNA			
Population	*N*	*h*	MNPD	SE	*N*	*h*	MNPD	SE
North								
Fali	43	16	1.859	1.086	41	26	8.233	3.896
Fulbe	27	22	3.288	1.745	34	26	6.734	3.255
Mandara	30	22	3.529	1.847	37	31	7.776	3.705
Podokwo	27	4	0.575	0.480	39	33	8.294	3.927
Tupuri	26	12	2.311	1.307	25	23	8.007	3.849
Uldeme	30	9	1.761	1.050	28	25	8.011	3.836
South								
Bamileke	50	19	2.535	1.387	48	36	8.108	3.830
Bakaka	49	12	1.900	1.102	50	36	9.821	4.572
Bassa	49	19	2.293	1.279	46	39	9.493	4.436
Ewondo	39	14	1.742	1.035	53	39	10.200	4.732

DISCUSSION

Implications of microsatellite variation for the genetic history of Cameroon

Our study of microsatellite variation provided some useful insights into the back-to-Africa migration hypothesized on the basis of the occurrence of the R1b1* (xR1b1b2) Y-chromosomal haplogroup at high frequencies in North Cameroon (Cruciani et al. 2002; see also Karafet et al. 2008 for nomenclature and Francalacci and Sanna, 2008 for a review).

The 29.2 allele at the D21S11 locus is rather rare in Africa, where it has been found at low frequency ($\leq$1%) in only 3 out of 31 populations so far studied (Guineans, Venda from South Africa, and Tunisians). It has also been reported at low frequencies in Europe (0.1–2.6%) and reaches the highest values in Asia (India, 8%) and in some Mexican indigenous populations (around 30%; http://alfred.med.yale.edu/; Rajeevan et al., 2003; Shepard et al., 2006). Previous work on Y-chromosome variation has reported a high frequency of the R1b1* (xR1b1b2) haplogroup of supposed Eurasian origin in northern Cameroon, with frequencies ranging from 7% (Tali) to 85% in Uldeme and 95% in Podokwo (data combined from Cruciani et al., 2002 and Wood et al., 2005). Taking the current 29.2/D21S11 distribution into account, it is possible that its occurrence among the Podokwo could be due to the same introgression event from non-African populations, which is responsible for the presence of the R1b1* (xR1b1b2) in that population. However, we cannot rule out the possibility that the 29.2/D21S11 distribution could be due to the persistence of a plesiomorphic character diluted in other populations by more recent events, or due to independent mutations (homoplasy), whose probability increases in microsatellites because of their high mutation rate.

The reduced haplogroup variation already noticeable in North Cameroon (see Cruciani et al., 2002), resulting from the high R1b1* (xR1b1b2) prevalence, was found to be associated to a noteworthy reduction of microsatellite diversity. In fact the HD and MNPD of some northern Cameroon populations (Podowko, Uldeme, and Tupuri) are the lowest observed in sub-Saharan Africa. The results of the search for undeclared family relationships suggest that the observed level of Y-chromosome diversity is probably the result of the specific demographic population history rather than the effect of a sampling bias. Interestingly, in a previous PCL-based study, we detected some signatures of genetic isolation in Podokwo and Uldeme through kinship analysis (Spedini et al., 1999, but see also MacEachern 2001 and Spedini et al., 2001). The present work provides new findings in this respect, by detecting robust signatures of isolation in the above-mentioned populations also for Y-chromosomal polymorphisms, including an extreme reduction of Y-chromosomal intra-population variation and increase in inter-population diversity. Interestingly, no signature of genetic drift was detected in the Podokwo and Uldeme using mtDNA polymorphisms (see below for further discussion).

A multi-perspective view of genetic variation

Grouping of populations on a geographic basis proved to be a simple and effective way of describing genetic diversity of Cameroon populations. In fact, all the genetic systems examined point to a greater variance of intra-population parameters and to a larger inter-population diversity for the northern part of the country. At the same time, all of the plots of genetic distances separate clearly northern and southern populations, with most of the genetic distances between them reaching statistical significance. These findings may be explained by the fact that the two areas have been shaped by distinct peopling processes, which have been more complex for North Cameroon. This is also mirrored by the linguistic diversity, since the southern populations under study belong to the Benue Congo sub-branch of the Niger Kordofanian phylum (Niger Congo branch), whereas those from northern Cameroon belong to two different linguistic phyla (Afro-Asiatic and the Niger-Kordofanian).

The discussion regarding the best predictor of genetic diversity is less straightforward. Previous and present analyses based on PCL data pointed to the role of language. Given the lack of knowledge of the exact molecular basis of PCL variation, we reanalyzed the Cameroon dataset using autosomal STRs. Unfortunately, the results neither clearly support nor contradict the role of language. Indeed, autosomal STRs provided the less clear and easily interpretable results among the genetic systems considered in this study. Further studies with broad autosomal SNP panels will probably provide a more useful test of the correlation between linguistic and genetic variation detected by using PCL data.

Concerning Y-chromosomal data, the effect of geography which is seen with AMOVA is lost in correlation analyses. On the other hand, geographic structure of mtDNA genetic diversity is detected by both types of analyses. The lack of robust signals of correlation between paternally inherited polymorphisms and language or geography contrasts with previous studies carried out on larger areas of sub-Saharan Africa (Poloni et al., 1997; Scozzari et al., 1999; Wood et al., 2005). Such a result should be considered in the light of the extreme reduction of Y-chromosome intra-populational diversity (and the correlated increase of diversity among populations) observed in northern Cameroon populations (see above).

On the other hand, the results obtained with mtDNA in our regional dataset are congruent with evidence obtained on a wider geographic scale (Salas et al., 2002) which have shown a higher percentage of variance between six main geographic groups (10.6%) compared

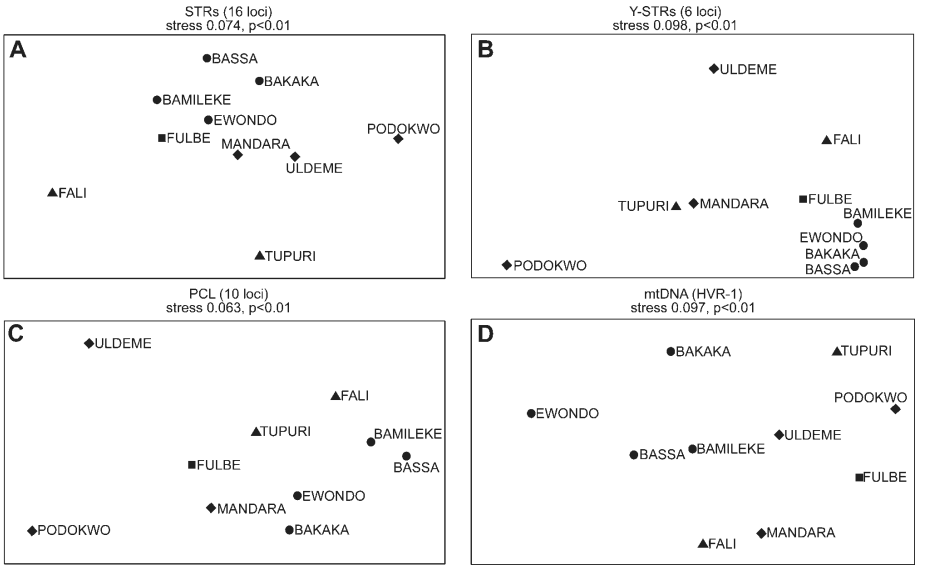


Fig. 2. Multidimensional scaling plot of the F_{st} genetic distances calculated for the autosomal microsatellite (STRs, **A**), Y-chromosome microsatellite (Y-STYRs, **B**), PCL (**C**) and mtDNA (**D**) in the 10 populations used for the comparison. Symbols used for population languages: diamond, Afro-Asiatic phylum and Chadic sub-branch; triangle, Niger-Kordofan phylum and Adamawa sub-branch; circle, Niger-Kordofan phylum and Benué-Congo sub-branch; rectangle, Niger-Kordofan phylum and West-Atlantic sub-branch.

TABLE 5. Analysis of molecular variance (AMOVA)

	Among groups				Among populations				Within populations			
	STRs	Y-STRs	PCL	mtDNA	STRs	Y-STRs	PCL	mtDNA	STRs	Y-STRs	PCL	mtDNA
1 group[a]					**1.1**	**34.8**	**1.0**	**2.9**	**98.9**	**65.2**	**99.0**	**97.1**
2 linguistic groups[b]	0.4	21.0	**0.6**	0.7	**1.0**	**22.0**	**0.9**	**2.6**	**98.7**	**57.0**	**98.5**	**96.7**
4 linguistic groups[c]	**0.4**	14.4	**0.5**	1.2	**0.8**	**23.0**	**0.7**	**2.0**	**98.8**	**62.7**	**98.9**	**96.7**
2 geographic groups[d]	0.2	**16.1**	0.2	**2.1**	**1.0**	**23.3**	**0.9**	**1.8**	**98.8**	**60.6**	**98.9**	**96.2**

[a] All populations were grouped together.
[b] Linguistic groups: Afro-Asiatic (Mada, Mandara, Podokwo and Uldeme) and Niger-Kordofan (Bakaka, Bamileke, Bassa, Ewondo, Fali, Fulbe and Tupuri).
[c] Linguistic groups: Afro-Asiatic (Mada, Mandara, Podokwo and Uldeme); Niger-Kordofan (Adamawa) (Fali, Tupuri), Niger-Kordofan (Benuè-Congo) (Bakaka, Bamileke, Bassa, Ewondo), Niger-Kordofan (West-Atlantic) (Fulbe).
[d] Geographic groups: North (Fali, Fulbe, Mandara, Podokwo, Tupuri and Uldeme) and South (Bakaka, Bamileke, Bassa, Ewondo).
Values are reported in percentage. Statistically significant values are reported in boldface ($P < 0.05$).

to that obtained dividing the database into the four main African linguistic groups (4.3%).

Another study has shown a weak correlation for the maternal transmitted marker both with geography and language (partial correlation values 0.17, $P = 0.035$ and 0.16, $P = 0.046$, respectively), and a significant positive correlation between Y-chromosome and linguistic variation (partial correlation values 0.33, $P = 0.001$) (Wood et al., 2005). This pattern has been interpreted as the result of a higher degree of female admixture and/or greater facility to adopt languages for females than males. Similarly, the pattern of mtDNA variation in Cameroon could have been shaped by a matrimonial mobility which has been less constrained by linguistic factors in females than in males.

The different behavior of the two unilinearly transmitted polymorphisms may be viewed in the light of a model we have previously proposed. It integrates demographic and genetic aspects and incorporates ethnographic knowledge, identifying differences between HGs (Hunter-Gatherers) and FPs (Food Producers) concerning direction of gene flow, extent of polygyny and respect of patrilocal-

TABLE 6. Correlation and partial correlation values calculated for the four genetic systems in the 10 populations used for comparison

	STRs		Y-STRs		PCL		mtDNA	
	r_s	P	r_s	P	r_s	P	r_s	P
Correlations								
gen × geo	0.118	0.283	0.213	0.057	0.223	0.067	0.578	**0.025**
gen × lin	0.220	0.113	0.422	**0.045**	0.510	**0.036**	0.108	0.159
geo × lin	0.319	**0.014**						
Partial correlations								
gen × geo (lin)	0.050	0.746	0.090	0.560	0.072	0.641	0.576	**0.000**
gen × lin (geo)	0.196	0.202	0.380	0.098	0.472	**0.001**	−0.097	0.529

Gen × geo (lin), partial correlations keeping the language constant; gen × lin (geo), partial correlations keeping the geography constant. Significant P values are reported in boldface.

ity as key factors for determining their diverse genetic structure (Destro-Bisol et al., 2004b). Applying the same approach to our population dataset, we obtained strikingly different N*v* estimates in FP populations from northern (25.82) and southern Cameroon (11.30), whose ranges do not overlap even when interval estimates produced by the jacknife procedure are considered (17.55–32.12 and 8.78–15.17, respectively; see Supporting Information Table S5). While the pattern observed in South Cameroon complies with what can be observed in other FPs (Destro-Bisol et al., 2004b), the ratios of mtDNA to Y-chromosome Nv of northern Cameroon populations can be distinguished due to their high values. Therefore, the results of the present study suggest the existence of a further heterogeneity among FPs regarding the ratio of female and male matrimonial mobility. How can this result be explained? There are two factors which could have exacerbated the discrepancy between mitochondrial and Y-chromosomal variation, potentially intensifying the pattern already observed among Food Producers in sub-Saharan Africa. It has been claimed that ethnographic and archaeological data suggest a high circulation of *Montagnard* women, producers of local pottery, among patrilocal communities in the northern region of Cameroon (MacEachern, 2001). According to the author, these cultural contacts might have been accompanied by a certain level of female gene flow and genetic admixture in northern Cameroon. On the other hand, previous studies have suggested a tendency towards isolation among *Montagnards* of northern Cameroon, related to physical barriers created by the mountainous environment (Boutrais, 1973). Our results on PCL and Y-chromosome variation are in line with this hypothesis (Spedini et al., 1999; present study).

CONCLUSIONS

This study provides new data and offers a synthesis of results on human genetic variation in Cameroon using data collected during a long-term anthropological study. By combining the information from genetic systems with different inheritance patterns and evolutionary rates with broader anthropological information, we have been able to better understand the genetic history of populations. We have also highlighted some specific aspects of particular groups, such as northern Cameroon populations, whose peculiarity is of particular importance even in the extremely varied genetic background of sub-Saharan populations. We have shown that the extremely complex pattern of genetic variation of Cameroon could in part be described by contrasting two geographic areas (corresponding to the northern and southern parts of the country) which differ substantially in environmental, biological and cultural terms. Moreover, the pattern of intra-regional variation has been highlighted, with a peculiar extreme reduction of Y-chromosome variation in some northern Cameroon populations.

Although this study does not address all the questions surrounding the complex processes of the peopling of Cameroon, we have identified some critical aspects, such as the need to test for potential sampling issues, a step that should be taken in consideration in future studies. Furthermore, we believe that further efforts to study genetic variation in Cameroon should involve the use of genomic approaches on population datasets adapted to the geographic and historical scale of the events under investigation. However, as our study illustrates, any attempt to interpret results and test hypotheses concerning the genetic history of human populations will always need to take cultural and historical aspects into proper consideration.

ACKNOWLEDGMENTS

We are greatly indebted to the blood donors, whose participation made this research possible. Cristian Capelli is an RCUK Academic Fellow.

LITERATURE CITED

Barreteau D, Breton R and Dieu M. 1984. Les langues. In: Boutrais J, editor. Le Nord du Cameroun. Des hommes, une région. Mémoires. Paris: ORSTOM. 102. p 159–180.

Boutrais J. 1973. La colonisation des plaines par les Montagnards au Nord du Cameroun. Monts Mandara. Paris: ORSTOM.

Butler JM. 2006. Genetics and genomics of core STR loci used in human identity testing. J. Forensic Sci 51:253–265.

Caglià A, Tofanelli S, Coia V, Boschi I, Pescarmona M, Spedini G, Pascali V, Paoli G, Destro-Bisol G. 2003. A study of Y-chromosome microsatellite variation in sub-Saharan Africa: a comparison between F(st) and R(st) genetic distances. Hum Biol 75:313–330.

Campbell MC, Tishkoff SA. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403–433.

Castrì L, Garagnani P, Useli A, Pettener D, Luiselli D. 2008. Kenyan crossroads: migration and gene flow in six ethnic groups from Eastern Africa. J Anthropol Sci 86:189–192.

Cerný V, Hájek M, Cmejla R, Brůzek J, Brdicka R. 2004. mtDNA sequences of Chadic-speaking populations from north-

ern Cameroon suggest their affinities with eastern Africa. Ann Hum Biol 31:554–569.

Cerný V, Salas A, Hájek M, Zaloudková M, and Brdicka R. 2007. A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet 71:433–452.

Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, Austerlitz F, Heyer E. 2007. From social to genetic structures in central Asia. Curr Biol 17:43–48.

Coia V, Caglià A, Arredi B, Donati F, Santos FR, Pandya A, Taglioli L, Paoli G, Pascali V, Spedini G, Destro-Bisol G, Tyler-Smith C. 2004. Binary and microsatellite polymorphisms of the Y-chromosome in the Mbenzele pygmies from the Central African Republic. Am J Hum Biol 16:57–67.

Coia V, Destro-Bisol G, Verginelli F, Battaggia C, Boschi I, Cruciani F, Spedini G, Comas D, Calafell F. 2005. Brief communication: mtDNA variation in North Cameroon: lack of Asian lineages and implications for back migration from Asia to sub-Saharan Africa. Am J Phys Anthropol 128:678–681.

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA. 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet 70:1197–1214.

David N. 1981. The archaeological background of Cameroonian history. In: Tardits C, editor. Colloques internationales du C.N.R.S. Paris. 551. p 80–97.

Destro-Bisol G, Boschi I, Caglià A, Tofanelli S, Pascali V, Paoli G, Spedini G. 2000. Microsatellite variation in Central Africa: an analysis of intrapopulational and interpopulational genetic diversity. Am J Phys Anthropol 112:319–337.

Destro-Bisol G, Coia V, Boschi I, Verginelli F, Cagliá A, Pascali V, Spedini G, Calafell F. 2004a. The analysis of variation of mtDNA hypervariable region 1 suggests that Eastern and Western Pygmies diverged before the Bantu expansion. Am Nat 163:212–226.

Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglià A, Tofanelli S, Spedini G, Capelli C. 2004b. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol 21:1673–1682.

Ehret C. 1984. Historical/linguistic evidence for early African food production. In: Clark JD, Brandt SA, editors. From hunters to farmers. Berkeley: University of California Press. p 26–35.

Excoffier L, Laval G, Schneider S. 2005. Arlequin ver. 3.1: an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491.

Francalacci P, Sanna D. 2008. History and geography of human Y-chromosome in Europe: a SNP perspective. J Anthropol Sci 86:59–89.

Gill P, Brinkmann B, d'Aloja E, Andersen J, Bar W, Carracedo A, Dupuy B, Eriksen B, Jangblad M, Johnsson V, Kloosterman AD, Lincoln P, Morling N, Rand S, Sabatier M, Scheithauer R, Schneider P, Vide MC. 1997. Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature. Forensic Sci Int 86:25–33.

Greenberg J. 1963. The languages of Africa. Bloomington: Indiana University Publications.

Greenberg J. 1980. Classification des langues d'Afrique. In: Ki Zerbo J, editor. Histoire générale de l'Afrique. Paris: UNESCO/ Jeune Afrique. p 321–346.

Harris DR. 1976. Traditional system of plant food production and the origins of agriculture in west Africa. In: Harlan J, De Wet J, Stemler A, editors. Origins of African plant domestication. The Hague: Mouton. p 311–356.

Jackson DA, Somers KM. 1989. Are probability estimates from the permutations model of Mantel's test stable? Can J Zool 67:766–769.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18:830–838.

Kayser M, de Knijff P, Dieltjes P, Krawczak M, Nagy M, Zerjal T, Pandya A, Tyler-Smith C, Roewer L. 1997. Applications of microsatellite-based Y-chromosome haplotyping. Electrophoresis 18:1602–1607.

Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V, Gehrig C, Bernini LF, Jespersen J, Bakker E, Roewer L, de Knijff P. 2001. An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. Am J Hum Genet 68:990–1018.

Konolalov DA, Manning C, Henshaw MT. 2004. KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. Mol Ecol Notes 4:779–782.

Kruskal JB. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1–27.

Lebeuf JP. 1981. Du rôle de l'archeology dans le connaissance du Cameroun. In: Tardits C, editor. Contribution de la récherche ethnologique a′ l'histoire des civilisations du Cameroun. Paris: CNRS. p 100–125.

MacEachern S. 2001. Montagnard ethnicity and genetic relations in Northern Cameroon: comment on "the peopling of sub-Saharan Africa: the case study of Cameroon," by G. Spedini, et al. Am J Phys Anthropol 114:357–360.

McIntosh SK, McIntosh RJ. 1983. Current direction in West African prehistory. Annu Rev Anthropol 12:215–258.

Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucl Acids Res 16:1215.

Mohammadou E. 1973. L'implantation des Peuls au Nord Cameroun. Paris: CNRS.

Moscetti A, Boschi I, Dobosz M, Destro-Bisol G, Pescarmona M, d'Aloja E, and Pascali VL. 1995. Fluorescence-based classification of microsatellites using a single-wavelength semiautomatic sequencer: genotype assignment and identity tests by analysis of comigrating peak profiles. Electrophoresis 16:1875–1880.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Pollard JH. 1977. Numerical and statistical techniques. London: Cambridge University Press. p 202–203.

Poloni ES, Semino O, Passarino G, Santachiara Benerecetti AS, Dupanloup I, Langaney A, Excoffier L. 1997. Human genetic affinities for Y-chromosome P49a,f/ TaqI haplotypes show strong correspondence with linguistics. Am J Hum Genet 61:1015–1035.

Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP, Yeh CC, Miller PL, Kidd KK. 2003. ALFRED: the ALlele FREquency Database—update. Nucl Acids Res 31:270–271.

Rakotomalala R. 2005. TANAGRA: un logiciel gratuit pour l'enseignement et la recherche. In: Zighed DA, Venturini G, editors. Actes de Extraction et gestion des connaissances. Paris RNTI-E-3. 2. p 697–702.

Relethford JH. 1985. Isolation by distance, linguistic similarity, and the genetic structure on Bougainville Island. Am J Phys Anthropol 66:317–326.

Relethford JH. 1990. Mantel: a microcomputer program for computing the Mantel probability between distance matrix elements. Oneonta, NY: Department of Anthropology, State University of New York, College at Oneonta.

Reynolds J. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–799.

Rousset F. 2008. Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. Mol Ecol Notes 8:103–106.

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo A. 2002. The making of the African mtDNA landscape. Am J Hum Genet 71:1082–1111.

Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellitto D, Arredi B, Destro-Bisol G, De Stefano G, Rickards O, Martinez-Labarga C, Modiano D, Biondi G, Moral P, Olckers A, Wallace DC, Novelletto A. 1999. Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. Am J Hum Genet 65:829–846.

Shepard EM, Herrera RJ. 2006. Genetic encapsulation among Near Eastern populations. J Hum Genet 51:467–476.

Spedini G, Destro-Bisol G, Mondovì S, Kaptuè L, Taglioli L, Paoli G. 1999. The peopling of sub-Saharan Africa: the case study of Cameroon. Am J Phys Anthropol 110:143–162.

Spedini G, Mondovì S, Paoli G, Destro-Bisol G. 2001. Biological and cultural contradictions? A reply to MacEachern. Am J Phys Anthropol 114:357–360.

StatSoft Italia 1997. Statistica per windows. http: //www. statsoft.com.

Sturrock K, Rocha J. 2000. A multidimensional scaling stress evaluation table. Field Methods 12:49–60.

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H, Hammer MF. 2005. Contrasting patterns of Y-chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur J Hum Genet 13:867–876.

Molecular Biology and Evolution

Research article

Signatures of the Preagricultural Peopling Processes in Sub-Saharan Africa as Revealed by the Phylogeography of Early Y Chromosome Lineages

Chiara Batini,†[,1,2] Gianmarco Ferri,[3] Giovanni Destro-Bisol,[2,4] Francesca Brisighelli,[4,5,6] Donata Luiselli,[7] Paula Sánchez-Diz,[6] Jorge Rocha,[8] Tatum Simonson,[9] Antonio Brehm,[10] Valeria Montano,[1,2] Nasr Eldin Elwali,[11,12] Gabriella Spedini,[2,4] María Eugenia D'Amato,[13] Natalie Myres,[14] Peter Ebbesen,[15] David Comas,[1] and Cristian Capelli*[,5]

[1]Institute of Evolutionary Biology, Universitat Pompeu Fabra-Institut de Biologia Evolutiva, Consejo Superior de Investigationes Cientificas, Barcelona, Spain

[2]Dipartimento di Biologia Ambientale, Sapienza Università di Roma, Rome, Italy

[3]Dipartimento Integrato di Servizi Diagnostici e di laboratorio e di Medicina Legale, cattedra di Medicina Legale, Università di Modena e Reggio Emilia, Modena, Italy

[4]Istituto Italiano di Antropologia, Rome, Italy

[5]Department of Zoology, University of Oxford, Oxford, United Kingdom

[6]Institute of Forensic Sciences Luis Concheiro, Genomics Medicine Group, University of Santiago de Compostela, Centre for Biomedical Network Research on Rare Diseases (CIBERER), Santiago de Compostela, A Coruña, Spain

[7]Dipartimento di Biologia Evoluzionistica Sperimentale, Unità di Antropologia, Università di Bologna, Bologna, Italy

[8]Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal

[9]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah School of Medicine

[10]Human Genetics Laboratory, University of Madeira, Campus of Penteada, Funchal, Portugal

[11]Department of Basic Sciences, College of Medicine, Al Imam Mohamed Bin Saud Islamic University, Riyadh, Kingdom of Saudi Arabia

[12]Department of Molecular Biology, National Cancer Institute (NCI-UG), University of Gezira, P.O. Box: 20, Wad Medani, Sudan

[13]Forensic DNA Lab, Department of Biotechnology, University of the Western Cape, Bellville, South Africa

[14]Sorenson Molecular Genealogy Foundation

[15]Laboratory for Stem Cell Research, University of Aalborg, Aalborg, Denmark

†Present address: Department of Genetics, University of Leicester, Leicester, United Kingdom

***Corresponding author:** E-mail: cristian.capelli@zoo.ox.ac.uk.

Associate editor: Beth Shapiro

Abstract

The study of Y chromosome variation has helped reconstruct demographic events associated with the spread of languages, agriculture, and pastoralism in sub-Saharan Africa, but little attention has been given to the early history of the continent. In order to overcome this lack of knowledge, we carried out a phylogeographic analysis of haplogroups A and B in a broad data set of sub-Saharan populations. These two lineages are particularly suitable for this objective because they are the two most deeply rooted branches of the Y chromosome genealogy. Their distribution is almost exclusively restricted to sub-Saharan Africa where their frequency peaks at 65% in groups of foragers. The combined high-resolution single nucleotide polymorphism analysis with short tandem repeats variation of their subclades reveals strong geographic and population structure for both haplogroups. This has allowed us to identify specific lineages related to regional preagricultural dynamics in different areas of sub-Saharan Africa. In addition, we observed signatures of relatively recent contact, both among Pygmies and between them and Khoisan speaker groups from southern Africa, thus contributing to the understanding of the complex evolutionary relationships among African hunter-gatherers. Finally, by revising the phylogeography of the very early human Y chromosome lineages, we have obtained support for the role of southern Africa as a sink, rather than a source, of the first migrations of modern humans from eastern and central parts of the continent. These results open new perspectives on the early history of *Homo sapiens* in Africa, with particular attention to areas of the continent where human fossil remains and archaeological data are scant.

Key words: Y chromosome, *Homo sapiens*, phylogeography, sub-Saharan Africa.

Introduction

In the last few decades, the analysis of genetic variation in human populations has increased exponentially and has provided significant insights on the history of our species (Destro-Bisol et al. 2010; Renfrew 2010). One of the most frequently replicated results has been the support of the

http://mbe.oxfordjournals.org/

"Recent Out of Africa" model, initially based on mitochondrial DNA (mtDNA; Cann et al. 1987) and later gaining support from other genomic regions (Underhill et al. 2001; Rosenberg et al. 2002; Li et al. 2008). Systematic investigation of the genetic diversity in African populations focusing on mtDNA (Salas et al. 2002; Behar et al. 2008), Y chromosomes (Underhill et al. 2001; Cruciani et al. 2002; Tishkoff et al. 2007), and autosomal regions (Tishkoff et al. 2009) has started to provide insights on African-specific demographic events. However, although mtDNA variation has been thoroughly investigated by detailed dissection of the most informative lineages (Salas et al. 2002; Gonder et al. 2007; Behar et al. 2008), and, more recently, autosomal variation has begun to be explored in detail (Tishkoff et al. 2009), such a level of resolution has been only partially applied to Y chromosome African haplogroups. Sub-Saharan African Y chromosome diversity is represented by five main haplogroups (hgs): A, B, E, J, and R (Underhill et al. 2001; Cruciani et al. 2002; Tishkoff et al. 2007). Hgs J and R are geographically restricted to eastern and central Africa, respectively, whereas hg E shows a wider continental distribution (see also Berniell-Lee et al. 2009; Cruciani et al. 2010). Despite the phylogeographic dissection of hg E is still ongoing, it has been suggested that this clade might be linked, at least in part, with the diffusion of agriculture and pastoralism in the continent during the last 4,000–5,000 years, as initially indicated by its parallel distribution to Bantu-speaking communities (Underhill et al. 2001; Henn et al. 2008). The other two lineages, A and B, represent the most basal branches within the human Y chromosome genealogy and are dispersed across different geographic areas and populations, with considerably high frequencies in hunter-gatherer populations. These hgs have been related to demographic dynamics that are independent to the recent introduction of practices for active food production mentioned above, thus suggesting an association with complex and potentially more ancient demographic events (Underhill et al. 2001; Cruciani et al. 2002; Tishkoff et al. 2007; Berniell-Lee et al. 2009).

In this work, we present a detailed phylogeographic dissection of hgs A and B in a broad data set of sub-Saharan populations, with the aim of providing new insights into the complex and poorly investigated dynamics that characterize the preagricultural history of sub-Saharan Africa, with special attention given to the relationships among Pygmy and Khoisan-speaking populations from southern Africa. In addition, we aim to contribute to the debate on the geographic origin of *Homo sapiens* in Africa by testing whether the male-specific signals of early human origins are retained only among communities from eastern Africa (as suggested by fossil remains and mtDNA; White et al. 2003; McDougall et al. 2005; Behar et al. 2008) or whether they can also be found within groups from southern Africa (as indicated by genome-wide scans and early Y chromosome analyses; Hammer et al. 2001; Semino et al. 2002; Hellenthal et al. 2008; Tishkoff et al. 2009).

Materials and Methods

Single Nucleotide Polymorphisms and Short Tandem Repeat Genotyping

A database of 641 chromosomes (supplementary table S1, Supplementary Material online) was generated by collecting previously published data, analyzing novel samples, and extending the molecular analysis of previously genotyped samples. All DNA samples were obtained from blood, buccal swabs, or saliva samples and collected from unrelated healthy individuals who gave the appropriate informed consent.

Samples were genotyped with different sets of markers (supplementary table S1, Supplementary Material online). Single nucleotide polymorphism (SNP) scoring was carried out using minisequencing multiplex reactions and direct sequencing. A total of 33 markers were selected within haplogroups A and B according to the most updated Y chromosome genealogy presented in Karafet et al. 2008. These were divided among four different single base extension (SBE) assays, here referred to as MAI, MAII, MB, and MB2b (see supplementary table S2, Supplementary Material online). Primers for multiplex PCR amplification were designed using Primer3Plus software (Untergasser et al. 2007) and are presented in supplementary tables S3 and S4 (Supplementary Material online). Self- and cross-compatibility among all primer pairs included in the same reaction were tested with the software Autodimer (see Web resources in Acknowledgments). Y chromosome specificity of each primer was tested using BlastN (basic local alignment search tool).

The Qiagen Multiplex PCR kit and conditions specified by the producer were applied with primer concentrations ranging between 0.15 and 0.8 μM. PCR products (1.5 μl) were cleaned using 1.5 μl of ExoSAP-IT (USB Corporation) for 15 min at 37 °C followed by 15 min at 80 °C.

Minisequencing SBE primers were selected using allele-specific primer extension tools in the National Institute of Standards and Technology (NIST) Online DNA Analysis tools Page (see Web resources in Acknowledgments), and nonspecific tails of different lengths were added to each in order to ensure complete capillary separation of SNaPshot products (supplementary tables S5 and S6, Supplementary Material online). The multiplex minisequencing assays were performed using 1 μl of purified product in a total volume of 5 μl using 2 μl of SNaPshot reaction mix (Applied Biosystems Carlsbad, CA) according to the SNaPshot Kit protocol. Fluorescently labeled dideoxy nucleotide triphosphates in excess were inactivated, and 1 μl of cleaned multiplex extension products were run on an ABI PRISM 3130 Genetic Analyzer. Allele calling was performed using GeneMapper software (v. 3.7; Applied Biosystems Carlsbad, CA, USA).

Direct sequencing was used to screen markers P108 and P114. Primers for amplification are reported in supplementary table S3 (Supplementary Material online). Amplification of MSY2 was carried out according to Bao et al. 2000.

Short tandem repeats (STR) genotyping was conducted using commercially available STR kits (Krenke et al. 2005;

Mulero et al. 2006) as well as multiplexes developed in-house (Beleza et al. 2003). All the samples included here were genotyped for ten STRs: DYS19, DYS389-I, DYS389-II (the allele reported in supplementary table S1, Supplementary Material online, has been obtained by subtracting the DYS389-I allele), DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439. A subset of the samples was tested for an additional five loci (DYS448, DYS456, DYS458, DYS635, and Y-GATA-H4). In the statistical analyses, specific loci (DYS385, DYS389-II, DYS390, DYS448, and DYS635) were excluded due to allelic homoplasy as reported in the NIST Y-STR Fact Sheets (see Web resources in Acknowledgments). Following this, eight STR loci were used in both phylogeographic and intralineage analyses in order to maintain broad population coverage.

Network Reconstruction and Diversity Estimation

Median-Joining networks (Bandelt et al. 1999) of both SNP and STR haplotypes were constructed using Network 4.5 (see Web resources in Acknowledgments). Weights were estimated using the inverse of the within-clade variances of individual STR loci. SNPs were weighted according to their hierarchical position in the genealogy identified in the present paper (see supplementary fig. S2*c* and *d*, Supplementary Material online). Within-hg diversity was investigated using Arlequin 3.0 (Excoffier et al. 2005). The variance was estimated as the within-locus mean allele variance averaged across all loci. Confidence intervals (CIs) were based on 10,000 resamplings performed across individuals. Samples showing missing data at any locus were not considered in the calculation of intralineage variation parameters.

Dating

The between- and within-lineage date estimates were obtained by using the model-free statistics average squared distance (ASD; Goldstein et al. 1995a, 1995b). An indication of the time of lineage split can be obtained using ASD calculated between lineages (ASD $= 2\mu T$; Goldstein et al. 1995a, 1995b). ASD is based on a strict single stepwise mutation model, and in the presence of multistep mutational events the squaring process is expected to heavily influence the distance estimation, corrupting the linearity with time. In order to take into account such occurrences and avoid the impact of multistep mutations, we calculated the expected ASD asymptotic value (Goldstein et al. 1995a) as an indication of the maximum expected ASD value per locus comparison. These values were used as locus-specific thresholds to identify and remove STR markers potentially showing between-lineage multistep mutational events. Mutation rate is a critical factor influencing the extension of ASD time-linearity. To control for this, we selected the set of eight markers among those available after multistep removal that showed the lowest mutation rate (based on the data presented on the Y-STR haplotype reference database (YHRD) webpage, release33; Willuweit et al. 2007; see also supplementary table S7, Supplementary Material online), for each interlineage comparison. In order to compare inter- and intralineage estimates, we used the same number of STRs (eight) for the within-lineage estimates (see below). ASD upper limit linearity with time can be estimated as described in Goldstein et al. (1995a). Simulations have shown that the expected values tend to overestimate the range of linearity and only provide a broad indication of the upper limit of ASD linearity with time (Goldstein et al. 1995a). We used these values as reference thresholds to ensure that all the between-lineage estimates reported in table 1*a* do not cross these boundaries. The starting set of markers comprised the 8 STRs used for Network analysis and diversity estimates and was extended to 11 by including DYS456, DYS458, and YGATA-H4 loci. Due to multistep correction, different sets of STRs were used (supplementary table S7, Supplementary Material online), and the average mutation rate was estimated using locus-specific values (YHRD, release33; Willuweit et al. 2007). The reported 95% CIs were estimated by averaging across the locus-specific upper and lower mutation rate estimates (YHRD, release 33; Willuweit et al. 2007). Given the limitation related to ASD saturation, some potentially interesting interlineage comparisons were beyond the available resolution dictated by the STRs we used, as, for example, the ASD between A and B clades, which is expected to give an estimate of the time to the most recent common ancestor (TMRCA) of the entire human Y chromosome genealogy. In order to provide an independent estimate for the TMRCA of a pair of lineages, we also used a Bayesian approach as described in Walsh (2001) and implemented in the software ASHEs (Tofanelli et al. 2009). In brief, this approach calculates the likelihood distribution of the TMRCA for each haplotype–haplotype comparison across n generations. In our analysis, the following parameters were used : $\lambda(1/\mathrm{Ne}) = 0.0002$ (Walsh 2001), 10,000 generations, and the same set of STRs/mutation rates as for the corresponding ASD calculations. The maximum likelihood estimations of the number of generations to the most recent common ancestor were collected for each run and the average of these values used to obtain an indication of lineage separation. To calculate the CI, the same procedure was repeated by using the average upper and lower estimates for the locus-specific mutation rate, which was also performed for the ASD-based estimates.

The TMRCA of a clade was estimated by calculating the ASD between all chromosomes in a lineage and the founder haplotype that we reconstructed by combining the modal alleles at single loci (Thomas et al. 1998). ASD estimated in this way has an expected value of μT, where μ is the average effective mutation rate at the loci and T is the separation time expressed in number of generations. This approach is expected to underestimate the age of the clade if the reconstructed founder haplotype differs from the true one. The 95% CIs were estimated using the software Ytime, based on a constant-size demographic model (Behar et al. 2003). The locus-specific mutation rate was estimated using data from the YHRD, release 33 (Willuweit et al. 2007). We focused on the same set of eight STRs used in the Network analyses. For estimates within hg A1, we removed the locus DYS438 due to its multistep behavior within this lineage and performed estimates on the

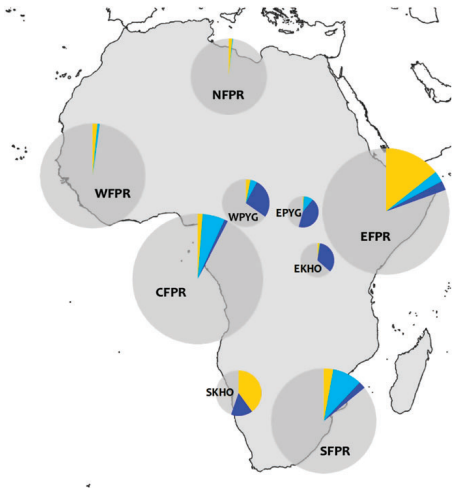


FIG. 1. Frequencies of haplogroups A (yellow), B2a (light blue), and B2b (dark blue) in Africa. For details on specific populations included in these groups, please refer to the column "Group code" in supplementary table S8 (Supplementary Material online). NFPR, northern food producers; WFPR, western food producers; WPYG, western Pygmies; CFPR, central food producers; EPYG, eastern Pygmies; EKHO, eastern Khoisan speakers; EFPR, eastern food producers; SKHO, southern Khoisan speakers; SFPR, southern food producers.

remaining seven STRs. It should be also noted that many of these lineages are particularly rare and that the within-clade variation might have been only partially surveyed, a condition that may divert current estimates toward the lower bound of the real genealogical depth (see Petraglia et al. 2010). For all estimates, a generation time of 31 years was used (Helgason et al. 2003). The average mutation rate used for the dating estimates ranges between 1.6 and 2.2 $\times$ 10^{-3} mutations per locus per generation depending of which set of STRS markers was used (supplementary table S7, Supplementary Material online). These values are not substantially different from other estimates based on pedigree data and are approximately two to three times faster than the more general and non-locus-specific "evolutionary" rate (6.9 $\times$ 10^{-4} mutations per locus per generation; Zhivotovsky et al. 2004; see also Ravid-Amir and Rosset 2010).

Results

Hg Distribution and Variation

We genotyped both novel and previously partially investigated samples and surveyed literature data for a total of ~10,000 males from more than 180 populations (supplementary table S8, Supplementary Material online), collecting data for 184 hg A and 457 hg B Y chromosomes (supplementary table S1, Supplementary Material online). Outside Africa, these clades have been sporadically found in Europe and the Americas, probably as a result of recent migrants (Semino et al. 2000; Luis et al. 2004; Capelli et al. 2006; Hammer et al. 2006; King et al. 2007). Hg A is rarely found in North, West, and Central Africa, whereas it is more frequent in the eastern and southern parts of the continent (fig. 1). Rare in both northern and western Africa, the distribution of hg B in the rest of the continent can be described by that of its two main subclades B2a and B2b (fig. 1). The former appears to be associated with food-producing communities and populations in contact with them, as also previously observed (Beleza et al. 2005; Berniell-Lee et al. 2009; Gomes et al. 2010), and it is present at low frequencies in all sub-Saharan areas. In contrast, B2b is mostly present in foraging communities in eastern and central Africa. The different geographic distributions of hgs A and B2b are mirrored at the population level (fig. 1 and supplementary table S8, Supplementary Material online). Little or no hg A is present in Pygmies and eastern African (EA) Khoisan speakers (for the use of the word Khoisan, issues with population classification in southern Africa, and the case of eastern Khoisan speakers, see Mitchell 2010), whereas B2b is commonly found in these populations. On the other hand, hg A is more frequent than B among southern African (SA) Khoisan speakers (~40%), with B2b representing ~16% of the Y chromosome types present in these populations (fig. 1 and supplementary table S8, Supplementary Material online).

Diversity indices are shown in table 2. Overall, hg A shows higher diversity than B and, within the latter, B2b is more variable than B2a. Network analysis based on eight STR haplotypes shows substantial phylogeographic patterns for A and B2b hgs (data not shown), whereas hg B2a reveals no clear population/geographic structure and a high level of reticulation, which is expected for lineages with a relatively short evolutionary history, associated with recent demographic expansions (table 1*b* and supplementary fig. S1, Supplementary Material online; see also Beleza et al. 2005; Berniell-Lee et al. 2009; Gomes et al. 2010). These results, together with the virtual absence of B2a in foraging populations, support our decision to focus the phylogeographic analysis on hgs A and B2b only, in order to address questions related to the early history of sub-Saharan Africa. The evolutionary relationships among haplotypes within these hgs, based on both SNPs and STRs, are shown in figure 2.

Of the A subclades, A1 is found only in western and central Africa, whereas A3b1 and A3b2 are southern and central/eastern African specific, respectively. Hg A2 is mostly represented by SA samples with only a few central African haplotypes (fig. 2*a* and *c*). Similarly, B2b1/B2b4a and B2b2 are geographically restricted to southern and eastern Africa, respectively, whereas B2b3, B2b4b, and B2b4* (as well as the previously undescribed MSY2* lineage; supplementary fig. S2*d*, Supplementary Material online) are specific to central Africa, albeit with few B2b4* SA haplotypes (fig. 2*b* and *c*). A prevalence of EA chromosomes is observed within B2b* together with considerable variation at the haplotype level, suggesting the possibility of yet undetected SNP-defined subclades within this group (fig. 2*b* and *c*). The geographically structured distribution within the B2b clade is shaped by the

Table 1. Between-Lineage (*a*) and Within-Lineage (*b*) TMRCA estimates based on ASD and maximum likelihood (ML).

	N	Years BP (95% CI)
(*a*) TMRCAs among lineages		
B2b2 versus B2b3 (ASD)	7 versus 7	10,695 (3,534–17,143)
B2b2 versus B2b3 (ML)		10,478 (6,882–16,523)
B2b2 versus B2b4b (ASD)	7 versus 6	14,322 (9,300–22,909)
B2b2 versus B2b4b (ML)		15,221 (10,013–23,932)
A2 SKHO versus WPYG (ASD)	5 versus 2	2,883 (1,891–4,619)
A2 SKHO versus WPYG (ML)		3,379 (2,201–5,363)
B2b4 SKHO versus WPYG (ASD)	3 versus 3	3,627 (2,356–5,766)
B2b4 SKHO versus WPYG (ML)		4,371 (2,821–6,913)
(*b*) TMRCAs within lineages (ASD with modal)		
A1-M31*	19	10,540 (4,185–23,684)
A1-M31 West Africa only*	12	8,091 (3,100–19,437)
A2-South	15	6,200 (2,232–14,198)
A3b1	22	10,261 (4,464–23,095)
A3b2	93	9,083 (3,720–20,274)
B2a	233	6,107 (2,263–14,012)
B2b3	10	1,984 (372–6,510)
B2b4b	11	713 (31–3,906)
B2b2	7	3,131 (868–8,990)

NOTE.—Generation time has been considered as 31 years (Helgason et al. 2003). Loci showing multistep mutational behavior were removed, and mutation rate per locus has been estimated as in the YHRD, release 33 (Willuweit et al. 2007; see supplementary table S7, Supplementary Material online, for details). For the clades indicated with (*), only seven STRs have been used for dating (see Materials and Methods for details). For population group abbreviations, refer to legend of figure 1 and supplementary table S8 (Supplementary Material online). *N*, number of chromosomes included in the calculation; BP, before present; SKHO, southern Khoisan speakers; WPYG, western Pygmies.

presence of population-specific lineages (fig. 2*b* and *c*). In fact, whereas B2b3, B2b4b, and B2b4* are almost exclusively found among western Pygmies, B2b2 and B2b1-B2b4a are found only in eastern Pygmies and SA Khoisan speakers, respectively. Similarly, the majority of the A3b1 and A2 types are found among SA Khoisan speakers, with hg A2 also present in western Pygmies (fig. 2*a* and *c*). Pygmies and SA Khoisan speakers also share evolutionarily closely related lineages within the B2b4 clade (fig. 2*b* and *c*; see also Wood et al. 2005).

A and B Genealogies

Our extensive survey of SNP variation in hgs A and B Y chromosomes enabled us to detect genealogical incompatibilities and propose some refinements within the recently proposed topology (see supplementary fig. S2, Supplementary Material online, for a comparison with the trees by Karafet et al. 2008). The PK1 marker, originally thought to be associated with the A2 lineage only, was found to cluster both A2 and A3 chromosomes. Similarly, new A2 lineages have been identified (supplementary fig. S2*c*, Supplementary Material online). M190 had been indicated as A3b-specific (Karafet et al. 2008); however, our analysis showed that it is derived in all A3 lineages. Within hg B, P7 appears to be basal to most of the B2b lineages and, within the P7-derived chromosomes, the MSY2 marker clusters lineages defined by M211, M115/M169, M30/M129 variants (supplementary fig. S2*d*, Supplementary Material online). The identification of two chromosomes derived at MSY2 and M30 (one of which is also derived for M129) but not for P7 suggests that this polymorphism might be prone to recurrent mutations (see fig. 2*c*). The physical proximity of P7 to P25 makes Y–Y gene conversion a possible explanation for this finding (see Adams et al. 2006). For simplicity, we have retained the same nomenclature as recently described in Karafet et al. 2008 (with the exception of MSY2*, see above), but renaming will be necessary as more data become available.

Discussion

Insights into the Male Genetic History of Sub-Saharan Hunter-Gatherers

Pygmy Groups

We dated the Eastern–Western Pygmy separation using the divergence between the B2b2 and B2b4b/B2b3 clades (table 1*a*). These estimates span similar time intervals and suggest a separation time of 10–15 thousand years ago (Kya), broadly overlapping with the generally more ancient estimates provided by mtDNA and autosomal data (Destro-Bisol et al. 2004; Patin et al. 2009; Batini et al.

Table 2. Diversity Indices for hg A and B, Including Subhaplogroups B2a and B2b, based on eight STRs (DYS19, DYS389I, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439).

Haplogroup	*N*	*k*/*N*	Haplotype Diversity (SD)	Variance (CI 2.5–97.5%)
A	180	0.589	0.988 (0.003)	1.099 (0.955–1.217)
B	443	0.400	0.987 (0.002)	0.562 (0.523–0.594)
B2a	233	0.373	0.965 (0.005)	0.294 (0.264–0.328)
B2b	184	0.451	0.980 (0.003)	0.743 (0.689–0.784)

NOTE.—Only samples with all the eight STRs available were included. *N*, number of chromosomes included in the calculation; *k*, number of different haplotypes; SD, standard deviation.

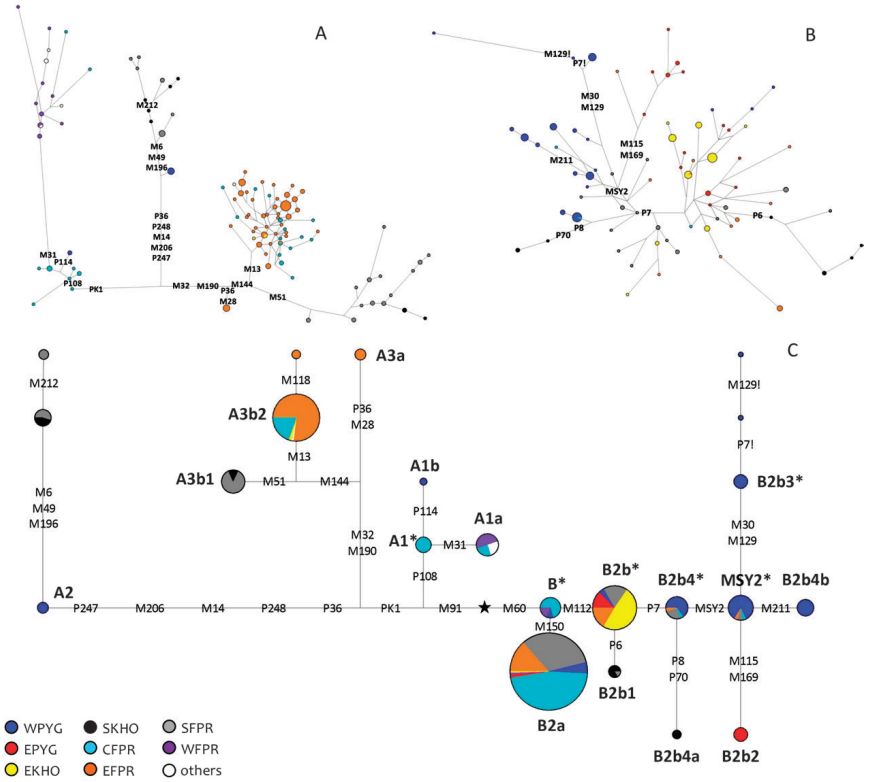


Fig. 2. Evolutionary relationships among A and B chromosomes. (*a*) Haplogroup A network, combined STR and SNP haplotypes; (*b*) haplogroup B2b network, combined STR and SNP haplotypes; (*c*) haplogroups A and B, SNP-based haplotypes. Haplotypes are colored according to the key in the figure, and circle size is proportional to the number of haplotypes, with the smallest representing $n = 1$. STR loci used in the present analysis are DYS19, DYS389-I, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439. The star represents the root of the Y chromosome tree as inferred from Karafet et al. 2008. For population group abbreviations, refer to legend of figure 1 and supplementary table S8 (Supplementary Material online). '!' indicates back-mutation.

2011). In particular, the youngest date suggested by B2b2 versus B2b3 (10.7 [3.5–17.1] Kya; 10.5 [6.8–16.5] Kya) might indicate post–Last Glacial Maximum (LGM; 19–26.5 Kya; Clark et al. 2009) male-mediated contacts between the two groups. This could account for the contrast between the lack of shared recent mitochondrial ancestry among Pygmy populations (Batini et al. 2011) and the quite intense post-LGM gene flow suggested by autosomal loci (Patin et al. 2009). However, the uncertainty related to STR choice and their time-linearity suggests that older scenarios might not be excluded (Busby G, Capelli C, personal communication). We also note that the within-clade diversity/antiquity is extremely reduced for these Pygmy-specific lineages, suggesting a bottleneck in the relatively recent demographic history of these groups, as it has been observed for other loci (see table 1*b*; Weiss and von Haeseler 1998; Excoffier and Schneider 1999; Patin et al. 2009; Batini et al. 2011).

Pygmies and San

We identified evolutionary links between western Pygmies and San in both A and B clades, developing the initial findings presented in Wood et al. (2005). Hg A2, found among SA Khoisan speakers at 25–45% (Wood et al. 2005; supplementary table S9, Supplementary Material online), was detected for the first time in the present work at nontrivial frequency (5%) among the Baka Pygmies from Cameroon and Gabon. On the other hand, B2b4 was present at 6–7% among Khoisan speakers but reached 45–67% in both Biaka and Baka Pygmies (Wood et al. 2005; supplementary table S9, Supplementary Material online). We dated the TMRCA among the western Pygmy- and San-specific

subclades of these two haplogroups to between 3 and 4 Kya (CI 1.9–4.6 and 2.2–5.4 Kya for A2; 2.3–5.8 and 2.8–6.9 Kya for B2b4; table 1*a*). It should be pointed out that the large number of mutations specific to the Khoisan A2 lineage (see fig. 2*a* and *c*) is probably the result of the SNP discovery process, which included Khoisan but not Western Pygmy A2 chromosomes (Underhill et al. 2001), thus making the use of STRs for dating the most obvious choice. Evidence for a Pygmy/San link has also been provided by recent genome-wide studies. In the work presented by Hellenthal et al. (2008), the first genetic link to emerge among human populations was indeed between the San and the western Pygmies. Furthermore, a shared ancestry between the San and the eastern Pygmies has been observed recently, and more generally, between the western Pygmies and the Hadza from Tanzania (Tishkoff et al. 2009), even though this has been interpreted as the result of a possibly more ancient common genetic background than the one suggested by our results. Intriguingly, the genetic link seems to be paralleled by the sharing of cultural traits such as those found in the rock art geometric designs produced by Pygmies from the Ituri forest and the Khoe-speaking groups from southern Africa (Smith 1995, 1997, 2006; Smith and Ouzman 2004). According to this model, Khoe-speaking pastoralists would have moved from an area in Central-South Africa bringing pastoralism into southern Africa before the Bantu dispersion in the region, having previously experienced cultural and genetic exchanges with central and EA populations (Henn et al. 2008; Rocha 2010).

Genetic Evidence for the Peopling of Sub-Saharan Africa before the Diffusion of Agriculture

West Africa

Haplogroup A in western Africa is represented only by the A1a lineage. The variation within this clade dates back to 10.5 (4.2–23.7) Kya and to 8 (3.1–19.4) Kya when only western African haplotypes are considered (see table 1*b*), which is in agreement with the archaeological and linguistic evidence related to the peopling of this region. The Ounanian culture has in fact been recorded in Mali as far back as 9–10 Kya (Clark 1980; Raimbault 1990; Mac Donald 1998), and the lithic and ceramic assemblages from Ounjougou date back to 12 Kya (Huysecom et al. 2004; Huysecom et al. 2009). Similarly, the origin of the early Niger-Congo Atlantic branch has been placed at least 8 Kya (Ehret 2000; Blench 2006). The detection of a specific genetic signal associated with early human presence in this area is of interest given the homogeneity between western and central African populations that has been observed so far for genome-wide analysis (Cruciani et al. 2002; Wood et al. 2005; Tishkoff et al. 2007; Li et al. 2008; Tishkoff et al. 2009).

South Africa

We dated variation in SA hgs A2 and A3b1 to 6.2 (2.2–14.1) Kya and 10.2 (4.4–23) Kya, respectively (table 1*b*). These dates do not extend beyond the LGM, which contrasts with the early human presence in southern Africa suggested by fossil and archaeological remains (McBrearty and Brooks 2000; White et al. 2003; Lewin and Foley 2004; McDougall et al. 2005; Marean et al. 2007). This could be possibly due to our partial population coverage, as suggested by extensive population surveys (Quintana-Murci et al. 2010; Marks S, Capelli C, unpublished data), as well as to past lineage extinctions (see Petraglia et al. 2010) that followed the significant demographic changes during the Marine Isotope Stage 3 (25–60 Kya) and the LGM (Mitchell 2008). Moreover, the possible limitation of available STRs in exploring events dating further back in time may also have had an effect (Busby G, Capelli C, personal communication). It is also worth considering the possibility that A2 and A3b1 retain signatures of two independent pre-Bantu dispersal events in the region. This scenario is also supported by the different geographic distribution of these two clades: A3b1 is present across all of southern Africa, whereas A2 is almost exclusively associated with populations in south-western Africa or those originally from this area (supplementary tables S1 and S9, Supplementary Material online; see also table 2 in De Filippo et al. 2010 and unpublished data from Lesotho and additional South African populations, where A3b1 but not A2 chromosomes were found—Marks S, Capelli C, personal communication). The A2 distribution broadly overlaps that of Khoe-speakers and could potentially represent a genetic signature of the contacts/migrations of the Khoe-speaking pastoralist societies from northern Botswana, southern Angola, and western Zambia area, ~2 Kya (see also above; Mitchell and Whitelaw 2005).

South-East Africa

Hg B2b4* chromosomes were present in the Mozambican samples, a lineage that is mainly shared with Baka Pygmies from Cameroon. The low frequency of these chromosomes in the SA and EA populations, together with the lack of appropriate evidence of a link among early inhabitants of these regions with western Pygmies, leaves the issue difficult to disentangle and calls for more detailed and focused investigation. In this sense, a scenario worth exploring could be based on the presence of this lineage in pre-Bantu populations already settled in the regions, which could either have been absorbed by the incoming agro-pastoralist groups (Sikora et al. 2010), or reflect the broader network of contacts around central-southern Africa (see above).

East Africa

The subclade A3b2 is present at high frequencies in EA populations, in particular among Nilo-Saharan speakers. Based on the analysis of this lineage in Uganda, Gomes et al. 2010 proposed its association with this linguistic phylum. Our estimates of A3b2 antiquity (9 Kya; CI 3.7–20.2 Kya) do not refute this hypothesis as they are broadly in agreement with the initial date for the spread of Nilo-Saharan phylum approximately between 12 and 18 Kya (Ehret 2000; Blench 2006).

B2a as a Marker of the Bantu Expansion?

Although B2a has not been investigated with the same resolution as the A and B2b hgs, our data support its

association with Bantu-speaking populations, as previously reported (see supplementary table S1, Supplementary Material online; Beleza et al. 2005; Berniell-Lee et al. 2009). Within-clade variation suggests a more recent origin for B2a than B2b, whereas network analysis did not reveal population-specific or geographically localized STR-based clusters (supplementary fig. S1, Supplementary Material online). However, the relatively deep within-clade dating (6.1 [2.2–14] Kya) suggests a scenario possibly pre-dating the diffusion of Bantu languages, in line with what has been observed for some subclades of hg E (Montano V, Destro-Bisol G, Comas D, personal communication). Deeper phylogenetic resolution within the B2a clade, coupled with additional population sampling, may help to clarify the demographic dynamics associated with its dispersal.

The Emergence of Modern Humans

Whereas the dissection of single Y-chromosomal clades or subclades has helped define the relationships between specific populations/groups, as well as reconstruct the demographic impact of migratory and cultural events, a wider and exhaustive phylogeographic analysis may indicate areas of the African continent where the extant human Y chromosome diversity first originated. Haplogroups A and B are ideal candidates for this task, given their distribution in Africa and the fact that they represent the earliest lineages to branch off within the Y chromosome genealogy. Previous analysis of the Y chromosome variation pointed to an SA/EA origin following the identification of hg A3b and, to a lower extent, B types in populations from these areas (Hammer et al. 2001; Semino et al. 2002). However, our results clearly indicate that A3b branched later within hg A, making it uninformative on the origin of the early human Y lineages. Hg A is divided into two branches: A1, represented by western and central African types, and A2-A3, containing SA and EA chromosomes, with a few from central Africa. Hg A2 is mostly composed of southern Africa types; however, an early branch in A2 is found in central Africa. Within hg A3, A3b1, the southern Africa clade, is a sister clade to A3b2, common in eastern Africa, whereas A3a is only found among EAs (fig. 2c). In hg B, B2a and B2b are two sister clades, whereas B*(xB2) aggregates a number of chromosomes from central Africa that were ancestral for the set of SNPs we tested. B2a has a very wide distribution and is mainly present in Bantu-speaking populations. Within hg B2b, B2b* contains samples from eastern, south-eastern, and central Africa, with P6-derived chromosomes from South Africa and P7 types mainly from hunter-gatherer populations from central, eastern, and southern Africa (see fig. 2c). These results seem to indicate that southern Africa was an early destination of ancient human migrations from other regions other than the original source, which fails to support the hypothesis presented in a recent large-scale study of autosomal loci (Tishkoff et al. 2009). With respect to the roles of eastern and central Africa, the data set presented here, although tentatively pointing toward a wide-scale preservation of ancient lineages in central Africa, is still compatible with a primary role for eastern Africa, in agreement with hypotheses generated from both mtDNA analysis and the study of the earliest *Homo sapiens* fossil remains (White et al. 2003; McDougall et al. 2005; Behar et al. 2008).

Concluding Remarks

Detailed phylogeographic analysis of human Y chromosome hgs A and B, combined with a large population survey and extensive sublineages characterization, has allowed us to gain new insights into the processes that shaped the pre-agricultural peopling of the African continent. Our results provide a male-specific perspective on some key aspects of the genetic history of sub-Saharan Africa and form the basis for future research.

We have shown evidence for further complexity in the evolutionary relationships among African hunter-gatherers. Phylogeographic analyses of mtDNA point to an ancient separation among ancestral populations, with limited or no subsequent gene flow after the split (see Salas et al. 2002; Destro-Bisol et al. 2004; Batini et al. 2007; Behar et al. 2008; Quintana-Murci et al. 2008; Batini et al. 2011). Conversely, the analysis of autosomal loci suggests a common, and possibly more recent, genetic background (see Tishkoff et al. 2009), with contrasting evidence concerning the reciprocal relationships among Pygmies and San (see Hellenthal et al. 2008; Li et al. 2008; Tishkoff et al. 2009), although this lacks a well-defined temporal context. Our extensive phylogeographic and dating approach has provided evidence for relatively recent contact both among Pygmies and between them and San groups from southern Africa. Our current estimates for the coalescent time between Eastern and Western Pygmy-specific Y chromosome clades (10–15 Kya) are compatible with post-LGM contact among the two groups, with evidence for recent bottlenecks in the demographic histories of the two groups (see also Patin et al. 2009; Batini et al. 2011). Otherwise, the very recent common ancestry detected among Western Pygmies and San (3–4 Kya) suggests that this could be the signature of Khoe-speaking pastoralist-mediated contact among the two groups, rather than resulting from retention of ancient traits.

Lastly, the peopling of sub-Saharan Africa has been studied from linguistic, archaeological, and genetic perspectives in the last decade, but its most ancient period is not yet well understood (see Campbell and Tishkoff 2010, Scheinfeldt et al. 2010). We have highlighted some signatures of preagricultural peopling undetected by previous research work. In fact, West, East, and South African populations show specific clades whose TMRCAs are compatible with a differentiation pre-dating the arrival of Bantu-speaking people and farming in the area. Intriguingly, even B2a, which has been mainly found in Bantu-speaking communities, has been dated (6 [2–14] Kya) before the supposed time of diffusion of Bantu languages. A novel link among Pygmy hunter-gatherers from west-central Africa and farmers from Mozambique has been identified, pointing to a shared genetic legacy between

these two geographically separate and anthropologically distinct population groups (see also Sikora et al. 2010).

Finally, our study contributes to the debate on the geographical origin of *Homo sapiens* in sub-Saharan Africa, providing evidence for the retention of early Y chromosome lineages in East and Central but not in Southern Africa. However, we note that the current absence of significant palaeo-anthropological investigation, together with the possibility of different fossil preservation conditions in central Africa, makes the extremely long human fossil record in eastern Africa inconclusive in solving this issue. The screening of Y-chromosomal variation at an increased level of resolution, combined with additional sampling from these regions, is expected to further elucidate the early steps of *Homo sapiens* in Africa.

Supplementary Material

Supplementary tables S1–S9 and figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Sergio Tofanelli and Davide Merlitti for giving access to early versions of the ASHEs software; Jim Wilson, Fabio Verginelli, and Renato Mariani-Costantini for providing samples and unpublished data; Peter Mitchell for helpful discussions on the African archaeological record; Marco Giorgi and Isabel Mendizabal for providing scripts used during data analysis; and Mònica Vallés, Stéphanie Plaza, and Roger Anglada (Universitat Pompeu Fabra) and Milena Alu' (Università di Modena e Reggio Emilia) for technical support. Finally, we would like to express our gratitude to all the people that have made this work possible by donating their DNA. The research presented was supported by the Dirección General de Investigación, Ministerio de Educación y Ciencia, Spain (CGL2007-61016), and Direcció General de Recerca, Generalitat de Catalunya (2009SGR1101). G.D.-B. and G.S. were supported by the University of Rome "La Sapienza" (C26A09EA9C/2009). J.R. was supported by the Fundação para a Ciência e a Tecnologia (PTDC/BIA-BDE/68999/2006). P.S.-D. is supported by the Isidro Parga Pondal program (Plan Galego de Investigación, Desenvolvemento e Innovación Tecnolóxica-INCITE [2006–2010] from Xunta de Galicia, Spain). C.C. is a Research Council UK Academic Fellow. C.B. and C.C. designed the research. C.B., G.F., D.C., and C.C. conceived and designed the experiments. G.D.-B., D.L., J.R., T.S., A.B., V.M., N.E.E., G.S., M.E.D.A., N.M., P.E., and D.C. provided the samples and part of the genotypings. C.B., G.F., F.B., and P.S.-D. performed the experiments. C.B. and C.C. analyzed the data. C.B. and C.C. wrote the paper with the contribution of G.D.-B. and D.C. All co-authors have reviewed the manuscript prior to submission. Web resources—Autodimer: http://cstl.nist.gov/; NIST Online DNA Analysis tools page: http://yellow.nist.gov:8444/dnaAnalysis/index.do; Y-STR Fact Sheets: http://www.cstl.nist.gov/strbase/ystr_fact.htm; Network 4.5: www.fluxus-engineering.com; ASHEs: http://ashes.codeplex.com/.

References

Adams SM, King TE, Bosch E, Jobling MA. 2006. The case of the unreliable SNP: recurrent back-mutation of Y-chromosomal marker P25 through gene conversion. *Forensic Sci Int.* 159:14–20.

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

Bao W, Zhu S, Pandya A, Zerjal T, Xu J, Shu Q, Du R, Yang H, Tyler-Smith C. 2000. MSY2: a slowly evolving minisatellite on the human Y chromosome which provides a useful polymorphic marker in Chinese populations. *Gene* 244:29–33.

Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, Comas D, Destro-Bisol G, Calafell F. 2007. Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. *Mol Phylogenet Evol.* 43:635–644.

Batini C, Lopes J, Behar DM, Calafell F, Jorde LB, van der Veen L, Quintana-Murci L, Spedini G, Destro-Bisol G, Comas D. 2011. Insights into the demographic history of African Pygmies from complete mitochondrial genomes. *Mol Biol Evol.* 28:1099–1110.

Behar DM, Thomas MG, Skorecki K, et al. (12 co-authors). 2003. Multiple origins of Ashkenazi Levites: Y chromosome evidence for both near eastern and European ancestries. *Am J Hum Genet.* 73:768–779.

Behar DM, Villems R, Soodyall H, et al. (15 co-authors). 2008. The dawn of human matrilineal diversity. *Am J Hum Genet.* 82:1130–1140.

Beleza S, Alves C, Gonzalez-Neira A, Lareu M, Amorim A, Carracedo A, Gusmao L. 2003. Extending STR markers in Y chromosome haplotypes. *Int J Legal Med.* 117:27–33.

Beleza S, Gusmao L, Amorim A, Carracedo A, Salas A. 2005. The genetic legacy of western Bantu migrations. *Hum Genet.* 117:366–375.

Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, van der Veen L, Hombert JM, Quintana-Murci L, Comas D. 2009. Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Mol Biol Evol.* 26:1581–1589.

Blench R. 2006. Archaeology, language, and the African past. Lanham (MD): AltaMira Press.

Campbell MC, Tishkoff SA. 2010. The evolution of human genetic and phenotypic variation in Africa. *Curr Biol.* 20:R166–R173.

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

Capelli C, Redhead N, Romano V, et al. (18 co-authors). 2006. Population structure in the Mediterranean basin: a Y chromosome perspective. *Ann Hum Genet.* 70:207–225.

Clark JD. 1980. Human populations and cultural adaptations in the Sahara and the Nile during prehistoric times. In: William MAJ, Faure H, editors. The Sahara and the Nile: quaternary environments and prehistoric occupation on Northern Africa. Rotterdam (The Netherlands): Balkema. p. 527–582

Clark PU, Dyke AS, Shakun JD, Carlson AE, Clark J, Wohlfarth B, Mitrovica JX, Hostetler SW, McCabe AM. 2009. The last glacial maximum. *Science* 325:710–714.

Cruciani F, Santolamazza P, Shen P, et al. (16 co-authors). 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet.* 70:1197–1214.

Cruciani F, Trombetta B, Sellitto D, Massaia A, Destro-Bisol G, Watson E, Beraud Colomb E, Dugoujon JM, Moral P, Scozzari R. 2010. Human Y chromosome haplogroup R-V88: a paternal genetic record of early mid Holocene trans-Saharan connections and the spread of Chadic languages. *Eur J Hum Genet.* 18:800–807.

De Filippo C, Heyn P, Barham L, Stoneking M, Pakendorf B. 2010. Genetic perspectives on forager-farmer interaction in the Luangwa valley of Zambia. *Am J Phys Anthropol.* 141:382–394.

Destro-Bisol G, Coia V, Boschi I, Verginelli F, Caglia A, Pascali V, Spedini G, Calafell F. 2004. The analysis of variation of mtDNA hypervariable region 1 suggests that eastern and western pygmies diverged before the Bantu expansion. *Am Nat.* 163:212–226.

Destro-Bisol G, Jobling MA, Rocha J, Novembre J, Richards MB, Mulligan C, Batini C, Manni F. 2010. Molecular anthropology in the genomic era. *J Anthropol Sci.* 88:93–112.

Ehret C. 2000. Language and history. In: Heine B, Nurse D, editors. African languages: an introduction. Cambridge: Cambridge University Press. p. 272–297

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

Excoffier L, Schneider S. 1999. Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc Natl Acad Sci U S A.* 96:10597–10602.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. 1995a. An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139:463–471.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. 1995b. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci U S A.* 92:6723–6727.

Gomes V, Sanchez-Diz P, Amorim A, Carracedo A, Gusmao L. 2010. Digging deeper into east African human Y chromosome lineages. *Hum Genet.* 127:603–613.

Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA. 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol.* 24:757–768.

Hammer MF, Chamberlain VF, Kearney VF, Stover D, Zhang G, Karafet T, Walsh B, Redd AJ. 2006. Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. *Forensic Sci Int.* 164:45–55.

Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL. 2001. Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol.* 18:1189–1203.

Helgason A, Hrafnkelsson B, Gulcher JR, Ward R, Stefansson K. 2003. A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am J Hum Genet.* 72:1370–1388.

Hellenthal G, Auton A, Falush D. 2008. Inferring human colonization history using a copying model. *PLoS Genet.* 4:e1000078.

Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, Cruciani F, Tishkoff SA, Mountain JL, Underhill PA. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc Natl Acad Sci U S A.* 105:10693–10698.

Huysecom E, Ozainne S, Raeli F, Ballouche A, Rasse M, Stokes S. 2004. Ounjougou (mali): a history of Holocene settlement at the southern edge of the Sahara. *Antiquity* 78:579–593.

Huysecom E, Rasse M, Lespez L, Neumann K, Fahmy A, Ballouche A, Ozainne S, Maggetti M, Tribolo C, Soriano S. 2009. The emergence of pottery in Africa during the tenth millennium cal BC: new evidence from Ounjougou (Mali). *Antiquity* 83:905–917.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 18:830–838.

King TE, Parkin EJ, Swinfield G, Cruciani F, Scozzari R, Rosa A, Lim SK, Xue Y, Tyler-Smith C, Jobling MA. 2007. Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. *Eur J Hum Genet.* 15:288–293.

Krenke BE, Viculis L, Richard ML, et al. (14 co-authors). 2005. Validation of male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. *Forensic Sci Int.* 151:111–124.

Lewin R, Foley R. 2004. Principles of human evolution. Oxford: Wiley-Blackwell.

Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioglu C, Roseman C, Underhill PA, Cavalli-Sforza LL, Herrera RJ. 2004. The Levant versus the horn of Africa: evidence for bidirectional corridors of human migrations. *Am J Hum Genet.* 74:532–544.

Mac Donald KC. 1998. Archaeology, language and the peopling of west Africa: a consideration of the evidence. In: Blench R, Spriggs M, editors. Archaeology and language II: correlating archaeological and linguistic hypotheses. London: Routledge. p. 33–66

Marean CW, Bar-Matthews M, Bernatchez J, Fisher E, Goldberg P, Herries AIR, Jacobs Z, Jerardino A, Karkanas P, Minichillo T. 2007. Early human use of marine resources and pigment in South Africa during the middle pleistocene. *Nature* 449:905–908.

McBrearty S, Brooks AS. 2000. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *J Hum Evol.* 39:453–563.

McDougall I, Brown FH, Fleagle JG. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* 433:733–736.

Mitchell P. 2008. Developing the archaeology of marine isotope stage 3. *S Afr Archaeol Soc Goodwin.* 10:52–65.

Mitchell P. 2010. Genetics and southern African prehistory: an archaeological view. *J Anthropol Sci.* 88:73–92.

Mitchell P, Whitelaw G. 2005. The archaeology of southernmost Africa from c. 2000 bp to the early 1800s: a review of recent research. *J Afr Hist.* 46:209–241.

Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK. 2006. Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. *J Forensic Sci.* 51:64–75.

Patin E, Laval G, Barreiro LB, et al. (15 co-authors). 2009. Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet.* 5:e1000448.

Petraglia MD, Haslam M, Fuller DQ, Boivin N, Clarkson C. 2010. Out of Africa: new hypotheses and evidence for the dispersal of Homo sapiens along the Indian Ocean rim. *Ann Hum Biol.* 37:288–311.

Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, Bormans C, van Helden PD, Hoal EG, Behar DM. 2010. Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. *Am J Hum Genet.* 86:611–620.

Quintana-Murci L, Quach H, Harmant C, et al. (23 co-authors). 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci USA.* 105:1596–1601.

Raimbault M. 1990. Pour une approche du néolithique du sahara malien. *Trav du LAPMO.* 1990:67–82.

Ravid-Amir O, Rosset S. 2010. Maximum likelihood estimation of locus-specific mutation rates in Y-chromosome short tandem repeats. *Bioinformatics* 26:i440–i445.

Renfrew C. 2010. Archaeogenetics—towards a 'new synthesis'? *Curr Biol.* 20:R162–R165.

Rocha J. 2010. Bantu-Khoisan interactions at the edge of the Bantu expansions: insights from southern Angola. *J Anthropol Sci.* 88:5–8.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science* 298:2381–2385.

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A. 2002. The making of the African mtDNA landscape. *Am J Hum Genet.* 71:1082–1111.

Scheinfeldt LB, Soi S, Tishkoff SA. 2010. Colloquium paper: working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. *Proc Natl Acad Sci U S A.* 107(Suppl 2):8931–8938.

Semino O, Passarino G, Oefner PJ, et al. (17 co-authors). 2000. The genetic legacy of paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science* 290:1155–1159.

Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet.* 70:265–268.

Sikora M, Laayouni H, Calafell F, Comas D, Bertranpetit J. 2010. A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. *Eur J Hum Genet.* 19:84–88.

Smith BW. 1995. Rock art in south-central Africa [PhD thesis]. Cambridge: Department of Archaeology, University of Cambridge.

Smith BW. 1997. Zambia's ancient rock art: the painting of Kasama. Oxford: Nuffield Press for the National Heritage Conservation Commission of Zambia.

Smith BW. 2006. Reading rock art and writing genetic history: regionalism, ethnicity and the rock art of southern Africa. In: Soodyall H, editor. The prehistory of Africa: tracing the lineage of modern man. Cape Town: Jonathan Ball Publishers. p. 76–96

Smith BW, Ouzman S. 2004. Taking stock: identifying Khoekhoen Herder rock art in southern Africa. *Curr Anthropol.* 45:499–526.

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB. 1998. Origins of old testament priests. *Nature* 394:138–140.

Tishkoff SA, Gonder MK, Henn BM, et al. (12 co-authors). 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol Biol Evol.* 24:2180–2195.

Tishkoff SA, Reed FA, Friedlaender FR, et al. (25 co-authors). 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035–1044.

Tofanelli S, Bertoncini S, Castri L, Luiselli D, Calafell F, Donati G, Paoli G. 2009. On the origins and admixture of Malagasy: new evidence from high-resolution analyses of paternal and maternal lineages. *Mol Biol Evol.* 26:2109–2124.

Underhill PA, Passarino G, Lin AA, Shen P, Mirazón Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet.* 65:43–62.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* 35:W71–W74.

Walsh B. 2001. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics* 158:897–912.

Weiss G, von Haeseler A. 1998. Inference of population history using a likelihood approach. *Genetics* 149:1539–1546.

White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, Howell FC. 2003. Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* 423:742–747.

Willuweit S, Roewer L. International Forensic Y Chromosome User Group. 2007. Y chromosome haplotype reference database (YHRD): update. *Forensic Sci Int Genet.* 1:83–87.

Wood ET, Stover DA, Ehret C, et al. (11 co-authors). 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet.* 13:867–876.

Zhivotovsky LA, Underhill PA, Cinnioğlu C, et al. (18 co-authors). 2004. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet.* 74:50–61.

3.2 Europe

Article 3. Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. *Am J Hum Genet.*

Article 4. Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. *Mol Phylogenet Evol.*

Article 5. Different historical demographic layers of modern-day Italy as revealed by a comprehensive analysis of mitochondrial DNA and Y-chromosome variation. *PlosOne.* Submitted

Article 6. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proc Biol Sci.*

Article 7. The Etruscan timeline: a recent Anatolian connection. *Eur J Hum Genet.*

Article 8. Moors and Saracens in Europe: estimating the medieval North African male legacy in southern Europe. *Eur J Hum Genet.*

Article 9. A 9-loci Y chromosome haplotype in three Italian populations. *Forensic Sci Int.*

Article 10. Patterns of Y-STR variation in Italy. *Forensic Science International Genetics.*

Article 11. Phylogenetic evidence for multiple independent duplication events at the DYS19 locus. *Forensic Sci Int Genet.*

Article 12. Allele frequencies of fifteen STRs in a representative sample of the Italian population. *Forensic Sci Int Genet.*

Article 13. Allele frequencies of the new European Standard Set (ESS) loci in the Italian population. *Forensic Sci Int Genet.*

REPORT

Mitochondrial Haplogroup U5b3: A Distant Echo of the Epipaleolithic in Italy and the Legacy of the Early Sardinians

Maria Pala,[1] Alessandro Achilli,[1,2] Anna Olivieri,[1] Baharak Hooshiar Kashani,[1] Ugo A. Perego,[1,3] Daria Sanna,[4] Ene Metspalu,[5] Kristiina Tambets,[5] Erika Tamm,[5] Matteo Accetturo,[1] Valeria Carossa,[1] Hovirag Lancioni,[2] Fausto Panara,[2] Bettina Zimmermann,[6] Gabriela Huber,[6] Nadia Al-Zahery,[1,7] Francesca Brisighelli,[8] Scott R. Woodward,[3] Paolo Francalacci,[4] Walther Parson,[6] Antonio Salas,[8] Doron M. Behar,[9] Richard Villems,[5] Ornella Semino,[1] Hans-Jürgen Bandelt,[10] and Antonio Torroni[1,*]

There are extensive data indicating that some glacial refuge zones of southern Europe (Franco-Cantabria, Balkans, and Ukraine) were major genetic sources for the human recolonization of the continent at the beginning of the Holocene. Intriguingly, there is no genetic evidence that the refuge area located in the Italian Peninsula contributed to this process. Here we show, through phylogeographic analyses of mitochondrial DNA (mtDNA) variation performed at the highest level of molecular resolution (52 entire mitochondrial genomes), that the most likely homeland for U5b3—a haplogroup present at a very low frequency across Europe—was the Italian Peninsula. In contrast to mtDNA haplogroups that expanded from other refugia, the Holocene expansion of haplogroup U5b3 toward the North was restricted by the Alps and occurred only along the Mediterranean coasts, mainly toward nearby Provence (southern France). From there, ~7,000–9,000 years ago, a subclade of this haplogroup moved to Sardinia, possibly as a result of the obsidian trade that linked the two regions, leaving a distinctive signature in the modern people of the island. This scenario strikingly matches the age, distribution, and postulated geographic source of a Sardinian Y chromosome haplogroup (I2a2-M26), a paradigmatic case in the European context of a founder event marking both female and male lineages.

According to the archaeological evidence, modern humans first entered Southwest Asia ~45–50 thousand years ago (kya), and Europe soon afterwards. The first modern Europeans came from the Levant,[1] but an almost concomitant arrival of related groups in European Russia from interior western Asia via the Caucasus or along the eastern coast of the Caspian Sea might have also occurred.[2,3] These findings are consistent with the proposal that modern Europeans might have developed from related groups living in several regional enclaves in the same broad geographic area of Southwest Asia[4] and the observation that mitochondrial DNA (mtDNA) variation in all modern European populations is completely embedded in the western Eurasian portion of the mtDNA phylogeny.[5]

Approximately 20 ky after the arrival of their ancestors from Southwest Asia, Europeans faced dramatic and rapid climatic changes, which peaked with the Last Glacial Maximum (LGM), centered at ~21 kya. Major gaps in the archaeological record reveal an abandonment of North and Central Europe[6] and a contraction of the human range to southern European regions that served as refugia.[7,8] The deglaciation sequence began with the Bølling warming about 15 kya but stabilized only at the end of the Younger Dryas cold snap 11.6 kya.[9–12] In the refugia, human genetic variation was affected by drift and founder events, but the effects were probably strongest for mtDNA and Y chromosome because of their uniparental transmission and reduced effective population size. Thus, pre-LGM mtDNA and Y chromosome haplotypes were differently preserved (or lost) in the various refugia, but at the same time new haplotypes arose as a result of the occurrence of novel mutations. When the climate improved and Paleolithic populations from European refugia repopulated the continent, some of these novel (or differently preserved) haplotypes also spread. They subsequently gave rise to new star-like haplogroups in the phylogeny, marking the expansion range from each refugium.

In the last 10 years, numerous studies have evaluated the distribution and extent of variation of haplogroups in European populations, and evidence of the overwhelming importance of the Franco-Cantabrian refugium for the repeopling of much of Western and Northern Europe at the beginning of the Holocene has been obtained by the age estimates and geographic distributions of mtDNA haplogroups H1, H3, V, and U5b1b.[5,13–21] Y chromosome haplogroups R1b1b2-M269, I1-M253, and I2b1-M223 support the important role of the Franco-Cantabrian refuge zone,[22–24] whereas other Y haplogroups (I2a1-M423 and

[1]Dipartimento di Genetica e Microbiologia, Università di Pavia, Pavia 27100, Italy; [2]Dipartimento di Biologia Cellulare e Ambientale, Università di Perugia, Perugia 06123, Italy; [3]Sorenson Molecular Genealogy Foundation, Salt Lake City, UT 84115, USA; [4]Dipartimento di Zoologia e Genetica Evoluzionistica, Università di Sassari, Sassari 07100, Italy; [5]Department of Evolutionary Biology, University of Tartu and Estonian Biocentre, Tartu 51010, Estonia; [6]Institute of Legal Medicine, Innsbruck Medical University, Innsbruck A-6020, Austria; [7]Department of Biotechnology, College of Science, University of Baghdad, Iraq; [8]Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses; and Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia 15782, Spain; [9]Molecular Medicine Laboratory, Rambam Health Care Campus, Haifa 31096, Israel; [10]Department of Mathematics, University of Hamburg, Hamburg 20146, Germany
*Correspondence: torroni@ipvgen.unipv.it
DOI 10.1016/j.ajhg.2009.05.004.

R1a1-M17) reveal that the Balkan and Ukrainian refuge zones were also major genetic sources[25–30] for the human recolonization of Europe.

In addition to the refugia mentioned above, another glacial refugium in Europe was the Italian Peninsula.[8] However, neither mtDNA nor Y chromosome studies have yet been able to identify haplogroups marking expansions from this area, thus suggesting a marginal role, if any, of this southern European area in the postglacial repeopling of Europe.

Haplogroup U5 is one of the most ancient mtDNA haplogroups found in Europe. It evolved mainly within Europe where it spread after being involved in the first settlement of the continent by modern humans.[4,31] Its phylogeny is characterized by two branches—U5a and U5b—which are common in most European populations,[19,32,33] with U5b further split into U5b1 and U5b2.[19] In 2006, a third uncommon branch, named U5b3, harboring the control-region motif 16169A-16192-16235-16270-16519-150 was detected only in Sardinia,[34] an island that remained unconnected with the mainland even when the sea level was lowest during the LGM[35] and that was probably the last of the large Mediterranean islands to be colonized by modern humans.[36]

To shed some light on the origin of haplogroup U5b3, we surveyed a wide range of European (and neighboring) populations for the presence of U5 mtDNAs lacking the diagnostic markers of haplogroups U5b1 and U5b2. For all subjects involved, an appropriate informed consent was obtained and institutional review boards at the Universities of Pavia, Tartu, Santiago de Compostela, at the Rambam Health Care Campus, and at the Sorenson Molecular Genealogy Foundation approved all procedures. Several mtDNAs with this feature were identified in Sardinia, in agreement with the presence of U5b3 in the island, but others were detected, at a very low frequency, also in other regions. With the exception of most mtDNAs from Sardinia, which harbored the previously described U5b3 control-region motif, almost all other U5 mtDNAs were characterized by a different but related control-region motif (16192-16270-16304-150).

To define the phylogenetic relationships between the U5b3 mtDNAs from Sardinia and the U5 mtDNAs with the related control-region motif, we completely sequenced a total of 43 mtDNAs and, together with nine previously published sequences (Table S1 available online), incorporated them in a phylogeny of haplogroup U5 (Figure 1). All sequences clustered in a U5 clade that is defined by a transition at np 7226 in the coding region—a mutation whose presence can be easily tested at the population level by a survey with the restriction enzyme DdeI. This clade splits into different minor subsets with a clear star-like pattern, including one branch that corresponds to the previously defined U5b3. This finding prompted us to revise the nomenclature and name the entire clade as U5b3, six of its main subsets as U5b3a-f, and the branch encompassing the Sardinian mtDNAs as U5b3a1a (Figure 1).

When all coding-region base substitutions are considered,[37] the average sequence divergence (±SE computed as in Saillard et al.[38]) of the 52 coding region sequences from the root of U5b3 is 2.19 ± 0.44 substitutions (Table 1)—a value virtually identical to those reported for haplogroups H1 (2.11 ± 0.23) and H3 (2.14 ± 0.28).[15] This finding indicates that U5b3 expanded at about the same time as H1 and H3. Table 1 reports also the average sequence divergences calculated by using only synonymous transitions.[39] Because the mutation rate of Mishmar et al.[37] is probably an overestimate, mainly caused by partial saturation of some synonymous mutations,[40] and that of Kivisild et al.[39] represents an underestimate,[41] we used the intermediate global coalescence time of modern human mtDNA recently proposed by Perego et al.[42] as a reference point for the internal calibration of both approaches. Accordingly, we converted the haplogroup sequence divergences into time estimates by using averaged time calibrations corresponding to 4610 years per coding-region substitution and 7650 years per synonymous transition (Table 1). With this approach, the coalescence time estimates for the entire U5b3 are between 10.1 ky and 8.1 ky.

To evaluate the distribution of haplogroup U5b3 in modern European (and neighboring) populations, we performed a survey of all U5 control-region motifs reported in almost 35,000 subjects from 81 population samples. For published and unpublished data sets for which only hypervariable segment I (HVS-I) data were available, U5 mtDNAs were affiliated within U5b3 when lacking 16189 or 16256 and harboring 16304. The presence or absence of the mutations 16169A, 16192, and 16235 was also considered. The results of this survey are reported in Table S2 and illustrated in the spatial distribution of Figure 2. Haplogroup U5b3 is virtually absent in the Near East (the single U5b3 mtDNA found in Iraq was completely sequenced) and North Africa and is rare in Europe where, with the exception of the frequency peak in Sardinians (3.8%), its frequency barely reaches 1% only in some Mediterranean populations.

Out of the 55 U5b3 mtDNAs detected in Sardinians, all but one (sequence n. 39 in Figure 1) are characterized by the diagnostic control-region motif of sub-haplogroup U5b3a1a, whose coalescence time estimate is between 4.6 ky and 6.3 ky (Figure 1 and Table 1). The phylogeny of Figure 1 includes 17 complete sequences belonging to this sub-haplogroup and, with the possible exception of sequence n. 22 that is classified as a generic "Italian" without regional details,[43] all are from Sardinia. A search for the U5b3a1a control-region motif in published data sets revealed only two matches (both 16169A-16192-16235-16270) outside Sardinia, one in Sicily[44] and one in Rome.[45] Details concerning the ancestry of the two subjects are not available, but the geographic proximity of Sardinia to the areas where they were detected makes it likely that they represent recent events of gene flow from the island. This would mean that U5b3a1a has arisen

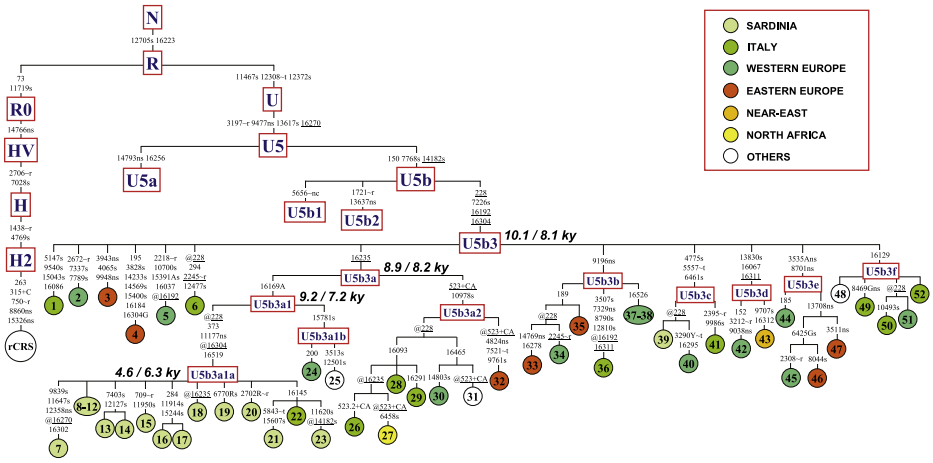


Figure 1. Detailed Tree of U5b3 in the Context of Haplogroup U5

The tree includes 52 complete mtDNA sequences and illustrates sub-haplogroup affiliations. The position of the revised Cambridge reference sequence (rCRS)[51] is indicated for reading off sequence motifs. MtDNAs were selected through a preliminary sequence analysis of the control region and an RFLP survey in order to include the widest possible range of internal variation of haplogroup U5b3. The sequencing procedure and phylogeny construction were performed as described elsewhere.[4,14,15] Sequences 1–9, 13–14, 18–19, 21, 24–52 are new while the others have been previously reported (Table S1). Mutations are shown on the branches; they are transitions unless a base is explicitly indicated. The prefix "@" designates reversions, whereas suffixes indicate: transversions (to A, G, C, or T), indels (+, d), gene locus (~t, tRNA; ~r, rRNA; ~nc, noncoding region outside of the control region), synonymous or nonsynonymous changes (s or ns), and heteroplasmies (R, Y). Recurrent mutations are underlined. The variation in number of Cs at np 309 was not included in the phylogeny: sequences 2, 4–5, 24, 30, 34–38, 47–49, 51–52 harbored 309+C, whereas sequence 50 harbored 309+CC. Additional information regarding each mtDNA is available on Table S1. Time estimates shown for clades are averaged distance (ρ) of each haplotype with respect to the respective root. The first value has been obtained by considering one coding-region substitution every 4610 years, whereas the second one assumes 7650 years per synonymous transition.

in situ in Sardinia after the arrival of an U5b3a1 founder mtDNA from somewhere else in Europe and that U5b3a1a affiliation is a marker of maternal Sardinian ancestry. The phylogeny of Figure 1 provides additional information concerning the entry time of the founder mtDNA; the upper limit is 9.2–7.2 ky (the age of U5b3a1 node), whereas the lower limit is 4.6–6.3 ky (the age of the U5b3a1a node), when the sub-haplogroup began to expand in Sardinia.

The phylogeny of Figure 1 also indicates a possible ancestral source for the founder(s) of the Sardinian U5b3a1a. The Sardinian-specific branch harbors a sister clade (U5b3a1b) formed by two sequences (n. 24 and 25): one from Languedoc, a region of southern France, and the other from a U.S. subject of undefined European ancestry. A search for the U5b3a1b control-region motif (16169A-16192-16235-16270-16304) was able to detect only one additional mtDNA from the southwestern (French-speaking) part of Switzerland,[46] matching such a motif. This preliminary observation suggests a stronger link between Sardinia and southern France than with other European regions, including continental Italy. Archaeological data from the period 5–10 kya show that the Monte Arci region of western Sardinia (Oristano province) was one of the four Mediterranean sources (together with the small islands of Palmarola, Lipari, and Pantelleria) of obsidian, the "black gold" of the Neolithic. In particular, a blooming trade of obsidian has been documented from Sardinia to other Mediterranean regions, including southern France. Moreover, it has been calculated that the obsidian employed in the Neolithic sites of the southern France was almost exclusively from a "single" Monte Arci subsource, suggesting not only a preferential link between French sites and Sardinia but also preferential transport mechanisms, different from those connecting Sardinia with other Mediterranean regions (Corsica and northern Italy) where this selection of specific subsources has not been detected.[47]

What about the ancestral homeland of the entire haplogroup U5b3? Its divergence is virtually identical to that reported for H1 and H3, thus indicating a population expansion at about the same time. Haplogroups H1 and H3 diffused from the Franco-Cantabrian refuge zone when climatic conditions improved;[15,18] therefore, it is

Table 1. Averaged Divergence of Relevant Nodes in the U5b3 Phylogeny of Figure 1

		All Coding-Region Base Substitutions				Only Synonymous Transitions			
Clade	No. of MtDNAs	ρ^a	σ^b	T^c(ya)	ΔT(ya)	ρ^a	σ^b	T^c(ya)	ΔT(ya)
U5b3	52	2.192	0.439	10,107	2,026	1.058	0.217	8,091	1,658
>U5b3a	26	1.923	0.744	8,865	3,429	1.077	0.357	8,238	2,729
> > U5b3a1	19	2.000	0.942	9,220	4,340	0.947	0.279	7,247	2,131
> > > U5b3a1a	17	1.000	0.294	4,610	1,356	0.824	0.276	6,300	2,111

[a] The average number of base substitutions in the mtDNA coding region (between positions 577 and 16023) from the root sequence type.
[b] Standard error calculated from an estimate of the genealogy.[38]
[c] Taking into account the limits of previous estimates reported by Mishmar et al.[37] for all coding-region base substitutions and by Kivisild et al.[39] for only synonymous transitions, we here employed a rate recently proposed by Perego et al.[42] With three decimal digits throughout, their rounded values were 5140 years per coding-region substitution and 6760 years per synonymous transition, respectively. The rho-estimated (average distance of the haplotypes of a clade from the respective root) human coalescence times are then 202 ky according to Mishmar et al.[37] and 160 ky according to Kivisild et al.[39] The postulated time obtained as the arithmetic mean of both estimates is ~181 ± 21 ky. Thus, ages estimated considering all the coding-region substitution have to be decreased by a factor of 181/202 ≈ 0.90, whereas the estimates based only on synonymous transitions have to be increased by a factor of 181/160 ≈ 1.13. Given that 5140 × 181/202 ≈ 4610 and 6760 × 181/160 ≈ 7650, we obtained the averaged calibrations.

possible that also the founder U5b3 sequence expanded from the same area and the three haplogroups were involved in the same demographic processes. However, there is also an alternative scenario: the expansion of U5b3 could have still occurred at the same time as H1 and H3 when climatic conditions in Europe changed, but from a distinct geographical source. With consideration to the modern range distribution of U5b3 (Figure 2), the only other potential candidate for the latter scenario is the glacial refuge in the Italian Peninsula.

To discriminate between the two possibilities, we measured the extent of U5b3 variation in different geographical areas by employing all available HVS-I (nps 16024–16365) data. A total of 152 U5b3 mtDNAs were detected, encompassing 40 HVS-I haplotypes, and their relationships are illustrated in the network of Figure 3. As expected, despite the frequency peak, Sardinians showed a very low haplotype diversity (H = 0.570), whereas much higher H values were observed in Italy (0.877) and Iberia (0.904) (Table 2), thus confirming that the relatively high frequency of U5b3 in Sardinia is the result of a founder event after the arrival on the island. Other indices such as nucleotide diversity and average number of nucleotide differences (Table 2), which are more informative than haplotype diversity because they take into account also the extent of diversity between haplotypes, not only confirm that Italy (0.717 and 2.45, respectively) and Iberia (0.645 and 2.21, respectively) are the European

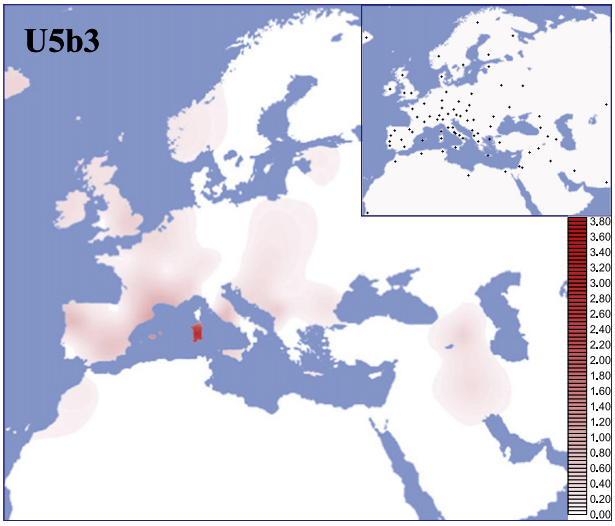


Figure 2. Spatial Frequency Distribution of Haplogroup U5b3 and Geographical Locations of Populations Surveyed
Populations and corresponding frequency values are listed in Table S2.

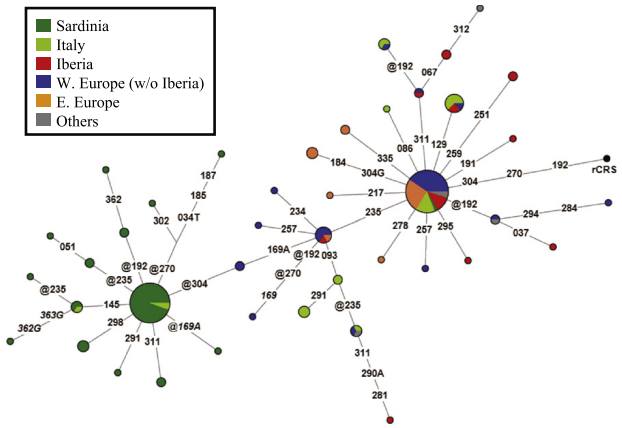


Figure 3. Median-Joining Network of HVS-I Haplotypes Observed in 152 U5b3 mtDNAs

Eighty-three mtDNAs are from the literature and a subset of these (N = 32) were not included in the population frequency table (Table S2) because population sample sizes were undefined. We constructed the tree by using the Network 4.510 software program (http://www.fluxus-engineering.com). The numbers (plus 16000) on the connecting branches refer to the revised reference sequence[51] and indicate mutations. These are transitions unless the base change is explicitly indicated; the prefix "@" designates reversions. Mutations in italics are most likely erroneous and were disregarded in the calculation of diversity measures. The size of each circle is proportional to the haplotype frequency and geographical origins are indicated by different colors. Fifty-five mtDNAs are from Sardinia; 23 are from Italy [continental Italy (N = 20) and Sicily (N = 3)]; 17 are from Iberia [Spain (N = 11), Portugal (N = 2) and Balearic Islands (N = 4)]; 33 are from Western Europe (excluding Iberia) [Belgium (N = 1), Denmark (N = 1), England (N = 6), France (N = 4), Germany (N = 3), Iceland (N = 3), Ireland (N = 3), Netherlands (N = 2), Norway (N = 1), Scotland (N = 6), Switzerland (N = 2), and Wales (N = 1)]; 19 are from Eastern Europe [Croatia (N = 4), Bosnia (N = 2), Bulgaria (N = 1), Crete (N = 1), Czech Republic (N = 4), Estonia (N = 1), Hungary (N = 2), Montenegro (N = 1), Poland (N = 1), and Slovakia (N = 2)]; and five are "Others" [Armenia (N = 1), Iraq (N = 1), Algeria (N = 1), and Morocco (N = 2)].

regions with the highest levels of U5b3 diversity but also reveal a peak in Italy, thus indicating continental Italy as the most likely focus of the U5b3 expansion.

Overall, the coalescence time of U5b3 (and those of the more common haplogroups H1 and H3) appears to indicate that the major post-LGM re-expansion phase in Europe was at the beginning of the Holocene (~11 kya) and not earlier. Whereas populations expanded geographically earlier during the warm phases of the Bølling-Allerød oscillations, the intermediate shorter-term cold phases and the Younger Dryas, in particular, led to retractions into the refugia again; it thus seems that in the Bølling-Allerød only some minor additional secondary refugia were created, which were too short-lived to leave discernible mutational marks in the mtDNA pools.

In contrast to the more common mtDNA haplogroups H1 and H3, however, the U5b3 diversity in modern Europe suggests that the glacial refuge located in the Italian Peninsula[8,48] rather than the Franco-Cantabrian refuge was the ancestral expansion source for haplogroup U5b3. Postglacial expansions of refugial populations from this area toward the North were restricted not only by cold phases but also by a geographical barrier—the Alps.[49] Thus, the ancestral U5b3 haplotype could have expanded (at a low frequency) outside the Italian Peninsula only along the coasts of the Tyrrhenian and Adriatic Seas, mainly toward the nearby Provence (southern France), and from there further west. The root of U5b3a1 originated probably in the Mediterranean coast of southern France and the same haplotype then went into Sardinia some 7–9 kya, possibly as a result of the obsidian trade that linked the two regions. There it expanded at the middle of the Neolithic, giving rise to an mtDNA clade (U5b3a1a) that distinctively marks the people of the island. Remarkably, the events leading to the arrival and expansion of this maternal lineage in Sardinia are not only supported but also magnified by data from

Table 2. Diversity of Haplogroup U5b3 MtDNAs in Different European Geographic Areas

Geographic Areas	No. of MtDNAs	No. of Haplotypes[a]	H[b]	π[c]	M[d]
Sardinia	55	13	0.570 ± 0.080	0.288 ± 0.062	0.986
Italy (continental Italy and Sicily)	23	9	0.877 ± 0.040	0.717 ± 0.093	2.451
Iberia (Spain, Portugal, and Balearic Islands)	17	10	0.904 ± 0.055	0.645 ± 0.117	2.206
Western Europe (w/o Iberia)	33	13	0.729 ± 0.081	0.409 ± 0.079	1.398
Eastern Europe	19	6	0.655 ± 0.111	0.315 ± 0.074	1.076

[a] HVS-I haplotypes (from np 16024 to np 16365).
[b] Haplotype diversity.
[c] Nucleotide diversity %.
[d] Average number of nucleotide differences.

male-specific lineages. Indeed ~37% of Sardinian Y chromosomes belong to haplogroup I2a2-M26,[50] a lineage rare outside Sardinia, whose age, distribution, and postulated geographic source (southern France)[28] strikingly match those of mtDNA haplogroup U5b3a1—a paradigmatic case of parallel founder events for both maternal and paternal lineages in the European context.

Supplemental Data

Supplemental Data include two tables and can be found with this article online at http://www.ajhg.org/.

Acknowledgments

This research received support from Fondazione Cassa di Risparmio di Foligno (to A.A. and F.P.), Fondazione Cassa di Risparmio di Perugia (to A.A. and F.P.), the European Union (European Regional Development Fund through the Centre of Excellence in Genomics), Estonian Biocentre, Estonian Science Foundation grants 7858 (to E.M.) and 6040 (to K.T.), the Swedish Collegium of Advanced Studies (to R.V.), Österreichische Forschungsförderungsgesellschaft, KIRAS Sicherheitsforschung DNATOX (to W.P.), Ministerio de Ciencia e Innovación (SAF2008-02971; to A.S.), Landau Network-Centro Volta (to N.A.-Z.), Progetti Ricerca Interesse Nazionale 2007 (Italian Ministry of the University) (to O.S. and A.T.), Ministero degli Affari Esteri (to O.S.), Compagnia di San Paolo (to O.S. and A.T.), and Fondazione Cariplo (to A.T.). We are grateful to all the donors for providing blood samples and to Cristian Capelli and the other people who helped collecting the samples.

Received: April 28, 2009
Revised: May 15, 2009
Accepted: May 15, 2009
Published online: June 4, 2009

Web Resources

The URLs for data presented herein are as follows:

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/
Network 4.510 software, http://www.fluxus-engineering.com

Accession Numbers

Previously unreported mtDNA sequences reported in this paper have been deposited in GenBank under accession numbers GQ129143–GQ129183.

References

1. Mellars, P. (2006). A new radiocarbon revolution and the dispersal of modern humans in Eurasia. Nature *439*, 931–935.
2. Goebel, T. (2007). The missing years for modern humans. Science *315*, 194–196.
3. Anikovich, M.V., Sinitsyn, A.A., Hoffecker, J.F., Holliday, V.T., Popov, V.V., Lisitsyn, S.N., Forman, S.L., Levkovskaya, G.M., Pospelova, G.A., Kuz'mina, I.E., et al. (2007). Early Upper Paleolithic in Eastern Europe and implications for the dispersal of modern humans. Science *315*, 194–196.
4. Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari, R., Cruciani, F., Behar, D.M., Dugoujon, J.-M., et al. (2006). The mtDNA legacy of the Levantine Early Upper Palaeolithic in Africa. Science *314*, 1767–1770.
5. Torroni, A., Achilli, A., Macaulay, V., Richards, M., and Bandelt, H.-J. (2006). Harvesting the fruit of the human mtDNA tree. Trends Genet. *22*, 339–345.
6. Street, M., and Terberger, T. (1999). The last Pleniglacial and the human settlement of Central Europe: New information from the Rhineland site of Wiesbaden-Igstadt. Antiquity *73*, 259–272.
7. Straus, L.G. (2005). The Upper Paleolithic of Cantabrian Spain. Evol. Anthropol. *14*, 145–158.
8. Banks, W.E., d'Errico, F., Peterson, A.T., Vanhaeren, M., Kageyama, M., Sepulchre, P., Ramstein, G., Jost, A., and Lunt, D. (2008). Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modelling. J. Archaeol. Sci. *35*, 481–491.
9. Alley, R.B., Meese, D.A., Shuman, C.A., Gow, A.J., Taylor, K.C., Grootes, P.M., White, J.W.C., Ram, M., Waddington, E.D., Mayewski, P.A., et al. (1993). Abrupt increase in Greenland snow accumulation at the end of the Younger Dryas event. Nature *362*, 527–529.
10. Björck, S., Kromer, B., Johnsen, S., Bennike, O., Hammarlund, D., Lemdahl, G., Possnert, G., Rasmussen, T.L., Wohlfarth, B., Hammer, C.U., and Spurk, M. (1996). Synchronized terrestrial-atmospheric deglacial records around the North Atlantic. Science *274*, 1155–1160.
11. Peteet, D. (2000). Sensitivity and rapidity of vegetational response to abrupt climate change. Proc. Natl. Acad. Sci. USA *97*, 1359–1361.
12. Alley, R.B., Marotzke, J., Nordhaus, W.D., Overpeck, J.T., Peteet, D.M., Pielke, R.A., Jr., Pierrehumbert, R.T., Rhines, P.B., Stocker, T.F., Talley, L.D., et al. (2003). Abrupt climate change. Science *299*, 2005–2010.
13. Torroni, A., Bandelt, H.-J., D'Urbano, L., Lahermo, P., Moral, P., Sellitto, D., Rengo, C., Forster, P., Savontaus, M.L., Bonné-Tamir, B., et al. (1998). MtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am. J. Hum. Genet. *62*, 1137–1152.
14. Torroni, A., Bandelt, H.-J., Macaulay, V., Richards, M., Cruciani, F., Rengo, C., Martinez-Cabrera, V., Villems, R., Kivisild, T., Metspalu, E., et al. (2001). A signal, from human mtDNA, of postglacial recolonization in Europe. Am. J. Hum. Genet. *69*, 844–852.
15. Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R., Cruciani, F., Zeviani, M., Briem, E., Carelli, V., et al. (2004). The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am. J. Hum. Genet. *75*, 910–918.
16. Forster, P. (2004). Ice ages and the mitochondrial DNA chronology of human dispersals: A review. Philos. Trans. R. Soc. Lond. B Biol. Sci. *359*, 255–264.
17. Gamble, C., Davies, W., Pettitt, P., and Richards, M. (2004). Climate change and evolving human diversity in Europe during the last glacial. Philos. Trans. R. Soc. Lond. B Biol. Sci. *359*, 243–253.
18. Loogväli, E.-L., Roostalu, U., Malyarchuk, B.A., Derenko, M.V., Kivisild, T., Metspalu, E., Tambets, K., Reidla, M., Tolk, H.-V., Parik, J., et al. (2004). Disuniting uniformity: A pied cladistic

canvas of mtDNA haplogroup H in Eurasia. Mol. Biol. Evol. *21*, 2012–2021.
19. Achilli, A., Rengo, C., Battaglia, V., Pala, M., Olivieri, A., Fornarino, S., Magri, C., Scozzari, R., Babudri, N., Santachiara-Benerecetti, A.S., et al. (2005). Saami and Berbers - an unexpected mitochondrial DNA link. Am. J. Hum. Genet. *76*, 883–886.
20. Pereira, L., Richards, M., Goios, A., Alonso, A., Albarrán, C., Garcia, O., Behar, D.M., Gölge, M., Hatina, J., Al-Gazali, L., et al. (2005). High resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res. *15*, 19–24.
21. Álvarez-Iglesias, V., Mosquera-Miguel, A., Cerezo, M., Quintáns, B., Zarrabeitia, M.T., Cuscó, I., Lareu, M.V., García, O., Pérez-Jurado, L., Carracedo, A., et al. (2009). New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS ONE *4*, e5112.
22. Semino, O., Passarino, G., Oefner, P.J., Lin, A.A., Arbuzova, S., Beckman, L.E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., et al. (2000). The genetic legacy of Paleolithic *Homo sapiens* in extant Europeans: A Y chromosome perspective. Science *290*, 1155–1159.
23. Wells, R.S., Yuldasheva, N., Ruzibakiev, R., Underhill, P.A., Evseeva, I., Blue-Smith, J., Jin, L., Su, B., Pitchappan, R., Shanmugalakshmi, S., et al. (2001). The Eurasian heartland: A continental perspective on Y-chromosome diversity. Proc. Natl. Acad. Sci. USA *98*, 10244–10249.
24. Zei, G., Lisa, A., Fiorani, O., Magri, C., Quintana-Murci, L., Semino, O., and Santachiara-Benerecetti, A.S. (2003). From surnames to the history of Y chromosomes: The Sardinian population as a paradigm. Eur. J. Hum. Genet. *11*, 802–807.
25. Passarino, G., Semino, O., Magri, C., Al-Zahery, N., Benuzzi, G., Quintana-Murci, L., Andellnovic, S., Bullc-Jakus, F., Liu, A., Arslan, A., et al. (2001). The 49a,f haplotype 11 is a new marker of the EU19 lineage that traces migrations from northern regions of the Black Sea. Hum. Immunol. *62*, 922–932.
26. Barać, L., Peričić, M., Martinović Klarić, I., Rootsi, S., Janićijevć, B., Kivisild, T., Parik, J., Rudan, I., Villems, R., and Rudan, P. (2003). Y chromosomal heritage of Croatian population and its island isolates. Eur. J. Hum. Genet. *11*, 535–542.
27. Cinnioğlu, C., King, R., Kivisild, T., Kalfoglu, E., Atasoy, S., Cavalleri, G.L., Lillie, A.S., Roseman, C.C., Lin, A.A., Prince, K., et al. (2004). Excavating Y-chromosome haplotype strata in Anatolia. Hum. Genet. *114*, 127–148.
28. Rootsi, S., Magri, C., Kivisild, T., Benuzzi, G., Help, H., Bermisheva, M., Kutuev, I., Barać, L., Peričić, M., Balanovsky, O., et al. (2004). Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europa. Am. J. Hum. Genet. *75*, 128–137.
29. Peričić, M., Barać Lauc, L., Martinović Klarić, I., Rootsi, S., Janićijevć, B., Rudan, I., Terzić, R., Čolak, I., Kvesić, A., Popović, D., et al. (2005). High-resolution phylogenetic analysis of southeastern Europe traces major episodes of paternal gene flow among Slavic populations. Mol. Biol. Evol. *22*, 1964–1975.
30. Battaglia, V., Fornarino, S., Al-Zahery, N., Olivieri, A., Pala, M., Myres, N.M., King, R.J., Rootsi, S., Marjanovic, D., Primorac, D., et al. (2008). Y-chromosomal evidence of the cultural diffusion of agriculture in southeast Europe. Eur. J. Hum. Genet., in press. Published online December 24, 2008. 10.1038/ejhg.2008.249.
31. Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., et al. (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. Am. J. Hum. Genet. *67*, 1251–1276.
32. Finnilä, S., Lehtonen, M.S., and Majamaa, K. (2001). Phylogenetic network for European mtDNA. Am. J. Hum. Genet. *68*, 1475–1484.
33. Tambets, K., Rootsi, S., Kivisild, T., Help, H., Serk, P., Loogväli, E.L., Tolk, H.V., Reidla, M., Metspalu, E., Pliss, L., et al. (2004). The western and eastern roots of the Saami - the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. Am. J. Hum. Genet. *74*, 661–682.
34. Fraumene, C., Belle, E.M.S., Castrì, L., Sanna, S., Mancosu, G., Cosso, M., Marras, F., Barbujani, G., Pirastu, M., and Angius, A. (2006). High resolution analysis and phylogenetic network construction using complete mtDNA sequences in Sardinian genetic isolates. Mol. Biol. Evol. *23*, 2101–2111.
35. Shackleton, J.C., van Andel, T.H., and Runnels, C.N. (1984). Coastal paleogeography of the central and western Mediterranean during the last 125,000 years and its archaeological implications. J. Field Archaeol. *11*, 307–314.
36. Sondaar, P.Y. (1998). Palaeolithic Sardinians: Paleontological evidence and methods. In Sardinian and Aegean chronology, M.S. Balmuth and R.H. Tykot, eds. (Oxford: Oxbow Books), pp. 45–51.
37. Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., et al. (2003). Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA *100*, 171–176.
38. Saillard, J., Forster, P., Lynnerup, N., Bandelt, H.-J., and Nørby, S. (2000). MtDNA variation among Greenland Eskimos: The edge of the Beringian expansion. Am. J. Hum. Genet. *67*, 718–726.
39. Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P.A., Scharfe, C., Torroni, A., et al. (2006). The role of selection in the evolution of human mitochondrial genomes. Genetics *172*, 373–387.
40. Tamm, E., Kivisild, T., Reidla, M., Metspalu, M., Smith, D.G., Mulligan, C.J., Bravi, C.M., Rickards, O., Martinez-Labarga, C., Khusnutdinova, E.K., et al. (2007). Beringian standstill and spread of Native American founders. PLoS ONE *2*, e829.
41. Achilli, A., Perego, U.A., Bravi, C.M., Coble, M.D., Kong, Q.-P., Woodward, S.R., Salas, A., Torroni, A., and Bandelt, H.-J. (2008). The phylogeny of the four pan-American mtDNA haplogroups: Implications for evolutionary and disease studies. PLoS ONE *3*, e1764.
42. Perego, U.A., Achilli, A., Angerhofer, N., Accetturo, M., Pala, M., Olivieri, A., Hooshiar Kashani, B., Ritchie, K.H., Scozzari, R., Kong, Q.-P., et al. (2009). Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr. Biol. *19*, 1–8.
43. Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. Nature *408*, 708–713.
44. Vona, G., Ghiani, M.-E., Calò, C.-M., Vacca, L., Memmì, M., and Varesi, L. (2001). Mitochondrial DNA sequence analysis in Sicily. Am. J. Hum. Biol. *13*, 576–589.
45. Turchi, C., Buscemi, L., Previderè, C., Grignani, P., Brandstätter, A., Achilli, A., Parson, W., Tagliabracci, A., and Ge.F.I. Group. (2008). Italian mitochondrial DNA

database: Results of a collaborative exercise and proficiency testing. Int. J. Legal Med. *122*, 199–204.
46. Dimo-Simonin, N., Grange, F., Taroni, F., Brandt-Casadevall, C., and Mangin, P. (2000). Forensic evaluation of mtDNA in a population from south west Switzerland. Int. J. Legal Med. *113*, 89–97.
47. Tykot, R.H. (2002). Chemical fingerprinting and source tracing of obsidian: The central Mediterranean trade in black gold. Acc. Chem. Res. *35*, 618–627.
48. Mussi, M. (2001). Earliest Italy. An overview of the Italian Palaeolithic and Mesolithic (New York: Plenum Publishing Company).
49. Taberlet, P., Fumagalli, L., Wust-Saucy, A.G., and Cosson, J.F. (1998). Comparative phylogeography and postglacial colonization routes in Europe. Mol. Ecol. *7*, 453–464.
50. Contu, D., Morelli, L., Santoni, F., Foster, J.W., Francalacci, P., and Cucca, F. (2008). Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: Inference for association scans. PLoS ONE *3*, e1430.
51. Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. *23*, 147.

Available online at www.sciencedirect.com
ScienceDirect
Molecular Phylogenetics and Evolution 44 (2007) 228–239
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic–Neolithic encounter

Cristian Capelli [a,b,*], Francesca Brisighelli [a], Francesca Scarnicci [a], Barbara Arredi [a,1], Alessandra Caglia' [a,2], Giuseppe Vetrugno [a], Sergio Tofanelli [c], Valerio Onofri [d], Adriano Tagliabracci [d], Giorgio Paoli [c], Vincenzo L. Pascali [a]

[a] *Institute of Legal Medicine, Universita' Cattolica del Sacro Cuore, Rome, Italy*
[b] *Department of Zoology, University of Oxford, South Park Road, Oxford OX1 3PS, UK*
[c] *Department of Biology, Anthropology Unit, University of Pisa, Pisa, Italy*
[d] *Institute of Legal Medicine, Universita' Politecnica delle Marche, Policlinico Torrette, Ancona, Italy*

Received 25 July 2006; revised 30 October 2006; accepted 21 November 2006
Available online 13 December 2006

Abstract

The Italian peninsula, given its geographical location in the middle of the Mediterranean basin, was involved in the process of the peopling of Europe since the very beginning, with first settlements dating to the Upper Paleolithic. Later on, the Neolithic revolution left clear evidence in the archeological record, with findings going back to 7000 B.C. We have investigated the demographic consequences of the agriculture revolution in this area by genotyping Y chromosome markers for almost 700 individuals from 12 different regions. Data analysis showed a non-random distribution of the observed genetic variation, with more than 70% of the Y chromosome diversity distributed along a North–South axis. While the Greek colonisation during classical time appears to have left no significant contribution, the results support a male demic diffusion model, even if population replacement was not complete and the degree of Neolithic admixture with Mesolithic inhabitants was different in different areas of Italy.

Keywords: Y chromosome; STRs; SNPs; Italian peninsula; Neolithic revolution

1. Introduction

The hominid presence in the Italian peninsula has been complex and extended in time. *Homo sapiens* probably made his first appearance in this area around 30–40 K years before present (ybp) (Cunliffe, 2001). Around 11,000 ybp in the Fertile Crescent new resources became available to humans in the means of domesticated crops and animals. The new technology was now able to support large communities and provided the resources for a demographic expansion (Cunliffe, 2001). Technology spread quite fast across the European peninsula, reaching the western fringes just 4000 years later (Ammerman and Cavalli-Sforza, 1984). The related demographic impact is still a matter of debate, but a consensus seems to have been reached on substantial Neolithic contribution in the Mediterranean area (Semino et al., 2000; Chikhi et al., 2001; Simoni et al., 2000). In Italy, Apulia, Calabria and Eastern Sicily were involved since the very beginning in this process as testified by the first Neolithic archaeological remains dating to around 9000 ybp. Farming technology appeared in Central Italy, on both sides

* Corresponding author. Fax: +44 (0) 1865 310447.
E-mail address: cristian.capelli@zoo.ox.ac.uk (C. Capelli).
[1] Present address: BMR Genomics, Padova, Italy.
[2] Present address: DAC Servizio Polizia Scientifica Sezione Genetica Forense, Via Tuscolana 1548-00173, Roma.

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.11.030

of the Apennines, and in the North East, in the Po and Adige Valleys, only 1000 years later. In the remaining areas, i.e. North and Central West Italy, farming technology arrived later, around 6.5K ybp and was characterised by a marked continuity with earlier Mesolithic groups. Indigenous communities in fact tended to select specific aspects of the new technology and integrate them with their existing ways of life (Cunliffe, 2001). This led to the presence of two well-defined farming groups in the peninsula: a North Italian–Tyrrhenian one and a South Italian-Adriatic one (Cunliffe, 2001). Distribution of genetic variation in Italy has only been investigated by a few studies. Barbujani and Sokal (1991) investigated the presence of genetic barriers within the peninsula by analysing classic polymorphisms. Interestingly, the zones of sharp changes in gene frequency were reflected at both linguistic and geographical level. MtDNA analysis revealed the presence of clines within the peninsula (Barbujani et al., 1995), as previously shown also by classical markers (Cavalli-Sforza et al., 1994). More recently, a Y chromosome investigation identified a single North–South major cline across the peninsula, (Di Giacomo et al., 2003), but pointed to local drift and founder effect as the main explanations for the observed distribution of genetic diversity.

Despite the archeological relevance and the fact that its position in the middle of the Mediterranean but connected to central Europe provides a preferential location for testing hypothesis related to the peopling of the continent itself, little efforts have been expended testing demographic scenarios shaping the currently observed genetic variation in Italian peninsula. Recently, Barbujani and Goldstein (2004) and Currat and Excoffier (2005), have both proposed two major models regarding the peopling of Europe: the demic diffusion model (DD) and the cultural diffusion model (CD), the major difference among the two being the demographic impact that Near Eastern farmers had on the European peninsula. The two models have different expectations in respect of the pattern of genetic variation and so concordance with observed results can be tested.

We investigated Italian Y chromosome variation by sampling a total of 699 chromosomes in 12 different areas along the peninsula. We genotyped these chromosomes using 10 microsatellites and 17 unique event polymorphisms, defining haplotypes (hpts) and haplogroups (hgs), respectively. In comparison to the most recent research in this area (Di Giacomo et al., 2003), we included a much larger set of markers, a more focused sampling scheme and larger average population size, a strategy that made possible a more comprehensive evaluation of the data. Results point to a distribution of genetic variation along a North–South axis and support the demic diffusion model. We discuss how this scenario could be explained by the admixture of two different groups: the Mesolithic original inhabitants of the peninsula and the incoming Neolithic farmers. Implications at continental levels are also discussed.

2. Materials and methods

2.1. Samples

Samples were selected according to father's place of origin. In order to have larger samples, we clustered samples smaller than 30 individuals with the closest sampling point within a 30-km radius. Collection was performed using buccal swabs and blood draws. DNA was extracted by a modified salting out protocol (Miller et al., 1988). Sample locations are described in Fig. 1. Sample sizes are in Table 1. TLB, CMA and SAP Y-STR haplotypes have been previously described (TL, MA and SA, respectively in Capelli et al., 2006a). For simplicity, as all samples except VLB were from the Italian peninsula, when referring to Italy we meant the geographic area of the peninsula, unless otherwise stated. European UEP and STR available data were included for comparison (Semino et al., 2000; Flores et al., 2004; Goncalves et al., 2005; Cinnioglu et al., 2004; Bosch et al., 2000, 2001; Capelli et al., 2006b; Roewer et al., 2005).

2.2. Genotypings

Microsatellite variation was investigated by the analysis of the following 10 microsatellites: DYS 388, 393, 392, 19, 390, 391, 389 I and II and 385—which is a double allele locus. PCR amplification was performed in two different multiplexes as previously described (Capelli et al., 2006a). Genotyped UEPs were as follows: M9, M17, M26, M35, M78, M81, M89, M173, M170, M172, M201, 92R7, 12f2, SRY10831, YAP, RPS4Y, tat. We first genotyped all the samples using a multiplex containing M173, M17, M172 and M170, following the protocol described in Capelli et al. (2006b). Then, we genotyped one after the other M35, M201, 12f2, M89, M9 (Capelli et al., 2006b). The M201 marker was genotyped as described in Flores et al. (2003). Samples ancestral at these markers, were additionally tested for YAP, RPS4Y and SRY10831 (Thomas et al., 1999; Capelli et al., 2001). Samples derived at M35 were additionally tested for M78 and M81. The two markers were co-amplified using following primers M78F 5′-tggctgtatgggtttctttga-3′ (label with Hex dye), M78R 5′-cttattttgaaatatttggaagagc-3′, M81F 5′-gcactatcatactcagctac acatctc-3′ and M81R 5′-ctataatattcagtacaatagtgtcgcac-3′ (label with Ned dye). The PCR products were then digested using Bsr B1and HpyCH 4 IV restriction enzymes and scored using a 373 AB automated sequencer. Samples derived at M9 but not derived at M173 were additionally tested for tat and 92R7 (Rosser et al., 2000; Capelli et al., 2006b). A subset of the samples was SNP genotyped as described in Onofri et al. (2006). Phylogenetic relationships between markers and nomenclature follow the Y Chromosome Consortium (YCC, 2002). In our investigation a number of Y chromosome STR duplications were found (see Supplementary material). In particular, we note that, differing from the previously described DYS19 duplicated

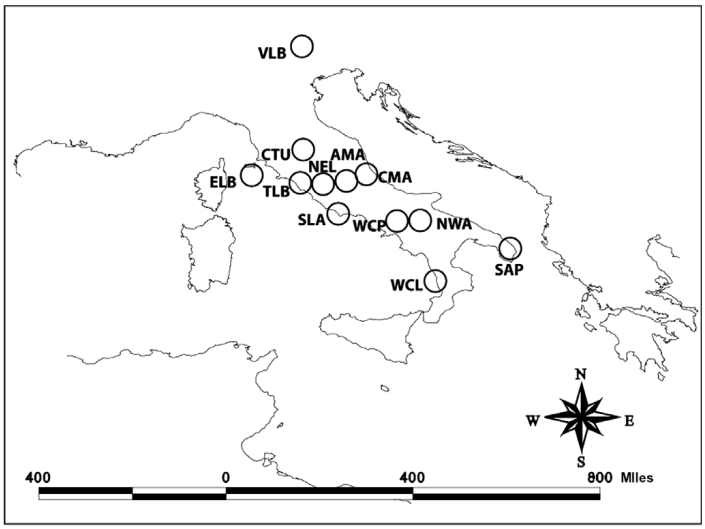


Fig. 1. Geographical location of the investigated Italian samples. Open circles showed the sample location areas. Sample codes are as follows: AMA, Apennine Marche; CMA, Central Marche; CTU, Central Tuscany; ELB, Elba Island (Tuscany); NEL, North–East Latium; NWA, North–West Apulia; SAP, South Apulia; SLA, South Latium; TLB, Tuscany–Latium border; VLB, Val Badia (Alto Adige); WCL, West Calabria; WCP, West Campania.

chromosomes, the ones here described were part of hg G instead of C3b (Zerjal et al., 2003; Nasidze et al., 2005). This will be presented in a more comprehensive way in a future publication. Haplotypes containing duplicated loci were not included in analysis based on STR variation.

2.3. Data analysis

Haplogroup frequencies were estimated by chromosome counting. European SNP and STRs data included Anatolia (Cinnioglu et al., 2004), Iberia (Goncalves et al., 2005; Flores et al., 2004; Bosch et al., 2000, 2001), Cyprus, Sicily, Sardinia, Malta (Capelli et al., 2006b), Albania (Semino et al., 2000; Pericic et al., 2004; Robino et al., 2002), Greece (Semino et al., 2000; Parreira et al., 2002). Principal component analysis was performed by POPSTR (Henry Harpending, personal communication). AMOVA analysis was performed by the Arlequin package (Schneider et al., 2000). Spatial correlation analysis was performed using hg frequencies using the AIDA software (Bertorelle and Barbujani, 1995). Geographic pattern of genetic variation were also investigated using Barrier 2.2 (Manni et al., 2004). Average squared distance (ASD) values were estimated by MICROSAT (Minch, 1996). The ADMIX software by Bertorelle and Excoffier (1998) was applied to estimate the Anatolian vs. West European contribution to the Italian samples (see Section 3) and performed on haplogroup frequencies.

3. Results

3.1. Italian Y chromosome haplogroup composition

The 699 chromosomes were SNP genotyped and clustered in 13 different haplogroups (Table 1) following the Y chromosome genealogy (YCC, 2002). Four haplogroups were not observed even if markers defining those were tested: A, C, N3 and Px(R). Haplogroups R1*(xR1a1), J2, G and E3b1 comprised more than 80% of the total chromosomes. The frequencies in the entire sample set were 40%, 20%, 11% and 10%, respectively. The remaining haplogroups had frequencies below 7% in the total sample, and never above 9% in a single population (except I*xI1b2 in the ELB sample at 16%). In all samples except two, WCL and SLA, R1*(xR1a1) was the most common haplogroup. Y-STRs were used to estimate intra-haplogroup diversity. Locus DYS385 has a duplicated allele pattern that can not be resolved assigning each allele to the corresponding locus. We thus decided to exclude DYS385 from STR variance estimation. Similarly, to avoid double estimation of locus variation, repeat number at locus DYS389 II was calculated by subtracting the number of repeats at DSY389 I. Intra-haplogroup STR diversity was estimated as the ASD (Slatkin, 1995; Goldstein et al., 1995a). The values for the four most common haplogroups are reported in Table 2.

Table 1
Italian samples studied

Location	Sample ID	Latitude	R1a1	R1* (xR1a1)		K* (xN3,P)		J2		J* (xJ2)		I1b2		I* (x1b2)		G		F (*xG,I,J,K)		E3b2		E3b1		E3b* (xE3b1,E3b2)		DE* (xE3b)		n
Val Badia (Alto Adige)	VLB	46.47		0.68	23	0.09	3	0.09	3					0.06	2	0.03	1					0.06	2					34
Central Tuscany	CTU	43.27	0.05	0.46	19	0.02	1	0.17	7	0.07	3			0.05	2	0.07	3					0.07	3	0.02	1			41
Central Marche	CMA	42.51	0.02	0.37	22			0.36	21	0.05	3			0.08	5	0.07	4					0.05	3					59
Elba Island (Tuscany)	ELB	42.48	0.01	0.53	50			0.08	8			0.04	4	0.16	15	0.11	10					0.06	6	0.01	1			95
Apennine Marche	AMA	42.31	0.07	0.33	9	0.04	1	0.22	6					0.04	1	0.15	4					0.07	2	0.07	2			27
Tuscany–Latium border	TLB	42.25	0.05	0.41	32	0.04	3	0.19	15					0.05	4	0.15	12					0.08	6	0.04	3			79
North–East Latium	NEL	42.21	0.02	0.38	21	0.02	1	0.15	8					0.05	3	0.13	7			0.02	1	0.22	12	0.02	1			55
South Latium	SLA	41.29	0.04	0.37	19	0.08	4	0.25	13			0.02	1	0.06	3	0.12	6					0.06	3					51
North–West Apulia	NWA	41.09	0.07	0.52	24	0.04	2	0.17	8	0.02	1					0.13	6			0.04	2							46
West Campania	WCP	41.07	0.02	0.29	24	0.08	7	0.17	17	0.01	1			0.03	2	0.11	8	0.04	3			0.20	14	0.03	2			71
West Calabria	WCL	39.21	0.02	0.32	18			0.35	20	0.04	2			0.02	1	0.11	6					0.16	9					57
			0.03	0.40	280	0.04	26	0.20	140	0.02	14	0.01	7	0.06	45	0.11	75	0.00	3	0.01	5	0.10	70	0.02	13	0.00	1	699

The table reports in order: the sample location, the sample code, the latitude of the sampling area and the haplogroup frequencies in percentage (left column) and number of individuals (right column). In the last raw are reported the sum of the hg frequencies and the related number of chromosomes.

3.2. Population structure

Population relationships among Italian samples were investigated by principal component analysis (data not shown). Thirty percent of the total variation was displayed by principal component 1 (PC1). Haplogroups R1*(xR1a1), J2, and E3b1 had the highest loading factors (+0.21, −0.14, −0.14, respectively) along this axis. We plotted PC1 sample values vs. sample latitudes obtaining a correlation between the two ($R^2 = 0.55$). In order to check for the presence of a geographic pattern of haplogroup distribution, we plotted the frequencies of the three hgs underlying PC1 in each population vs. the sample latitude. The regression lines are shown in Fig. 2A. R1*(xR1a1) frequencies tend to increase linearly moving north, while E3b1 and J2 decrease along the same direction. We additionally investigated the distribution of genetic variation as expressed by ASD within each of the three haplogroups (Table 2) and plotted these values vs. the sample latitudes, as shown in Fig. 2B. The regression line for R1*(xR1a1) showed very little correlation between ASD and latitude, while both E3b1 and J2 showed a linear decrease for ASD with increasing latitude. We additionally tested how Italian samples are related to each other within a European context. Principal component analysis was performed including European data (Semino et al., 2000; Flores et al., 2004; Goncalves et al., 2005; Cinnioglu et al., 2004; Bosch et al., 2000, 2001; Capelli et al., 2006b) (Fig. 3). European populations are distributed along axis one following a W–SE direction, a result already shown by others (Semino et al., 2000; Rosser et al., 2000). This pattern has been interpreted as the result of the admixture of Neolithic Near Eastern farmers with the European Mesolithic groups following the demographic expansion that was likely associated with the development of agricultural technology (Semino et al., 2000; Rosser et al., 2000). When included, the Italian samples do not cluster all together. Northern and Southern samples in fact tend to show along axis one negative and positive values respectively, suggesting a closer affinity to Western European populations for the former and to South–East and South–Central Europe for the latter, with few exceptions (see below) (Fig. 3). In order to test this interpretation, we performed the analysis of the distribution of genetic variation (AMOVA) using available Y-STR European data (Roewer et al., 2005; Cinnioglu et al., 2004; Parreira et al., 2002; Robino et al., 2002; Pericic et al., 2004), including the Italian samples. When South Italian samples were clustered with SouthEast and South-Central European samples and the Northern groups with West Europe, the percentage of between group variation was 6.04% and the within group 1.70% (Φ_{sc} 0.018, $P \ll 0.01$; Φ_{ct} 0.06, $P \ll 0.01$), while the exchange of the Italian groups decreased the between groups variation and increased the within group value (3.77% and 2.67% respectively, Φ_{sc} 0.027, $P \ll 0.01$; Φ_{ct} 0.037, $P \ll 0.01$).

The evidence for differential clustering tendency for North and South samples was confirmed when AMOVA

Table 2
ASD (average squared distance) values and associated Standard Error (SE) for the four most frequent hgs in each Italian sample

Code	Latitude	R1*(xR1a1)	SE	J2	SE	G	SE	E3b1	SE
VLB	46.47	0.467	0.148	0.278	0.073	0.000	0.000	0.125	0.075
CTU	43.27	0.486	0.140	0.612	0.135	0.611	0.198	0.111	0.068
CMA	42.51	0.298	0.055	0.629	0.263	0.167	0.075	0.167	0.075
ELB	42.48	0.544	0.057	0.758	0.157	0.615	0.222	0.257	0.059
AMA	42.31	0.623	0.166	0.583	0.149	0.406	0.160	0.125	0.078
TLB	42.25	0.701	0.121	0.656	0.232	0.707	0.135	0.347	0.090
NEL	42.21	0.529	0.140	0.410	0.093	0.750	0.334	0.297	0.176
SLA	41.29	0.414	0.073	0.837	0.280	0.486	0.153	0.167	0.073
NWA	41.09	0.297	0.058	0.840	0.450	0.120	0.072	—	—
WCP	41.07	0.527	0.106	0.892	0.233	0.667	0.226	0.210	0.107
SAP	39.50	0.533	0.126	0.979	0.185	0.383	0.091	0.457	0.140
WCL	39.21	0.500	0.089	0.754	0.201	0.472	0.123	1.062	0.319

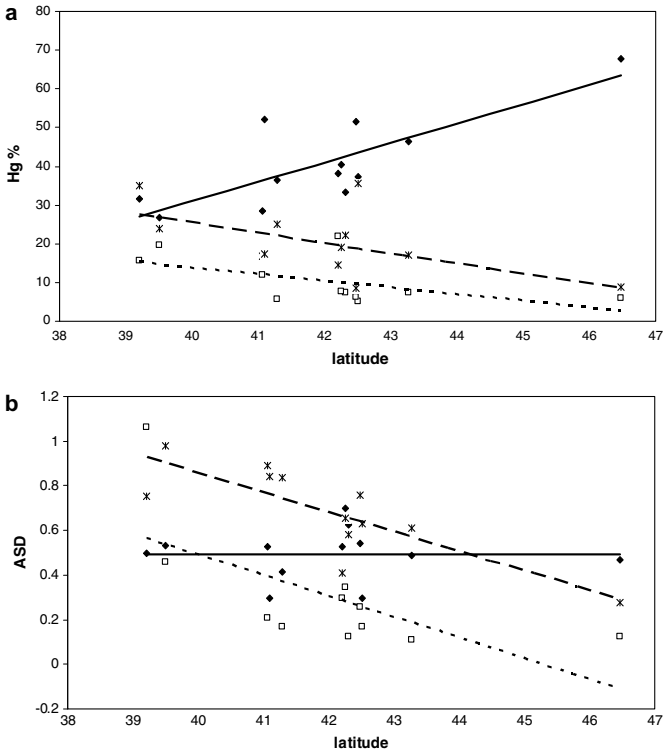


Fig. 2. Regression lines of the hg frequencies/average squared distance (ASD) versus latitude. (a) The regression lines of the latitude values of the selected samples (*X* axis) are plotted versus hg frequencies (*Y* axis); thick line, hg R1(*xR1a1), dashed line, hg J2; dotted line hg E3b1. E3b1 hg was absent from the NWA sample: this data point was not considered; (b) the latitude values (*X* axis) are plotted vs. ASD (*Y* axis) within selected hg; thick line, hg R1*(xR1a1); dashed line, hg J2; dotted line, hg E3b1. The E3b1 hg was absent from the NWA sample: this data point was not considered.

was conducted on the Italian samples only. Samples were divided in two groups, broadly defined as North and South, on the basis of their latitude (above and below 41.50° latitude, Table 3). This grouping displayed a negative value for the between groups variation (−0.19%, Φ_{ct} −0.001, $P = 0.46$), with a within group variation of

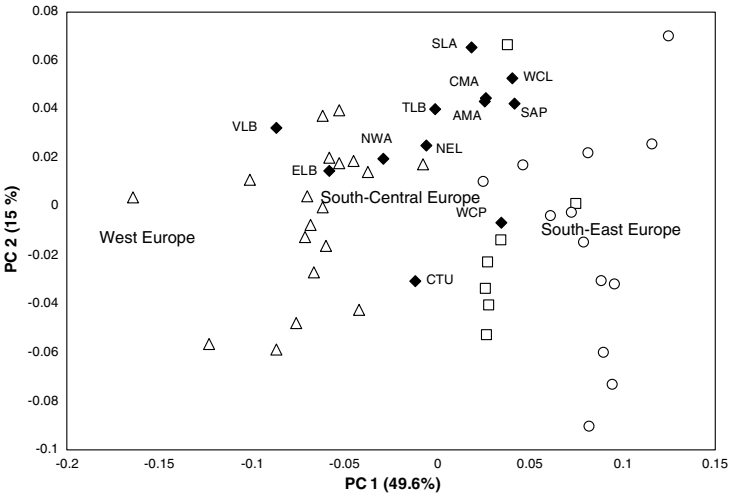


Fig. 3. Plot of the principal component analysis conducted on the haplogroup frequencies of the Italian and European samples; Italian samples investigated in this study are indicated by black diamonds, codes as in Table 1. Open circles, South–East European samples; open squares, South–Central European samples; open triangle, West European samples.

Table 3
Grouping schemes followed for the AMOVA analyses (see text)

	Latitude		PC	
	North	South	North	South
VLB	x		x	
CTU	x		x	
CMA	x			x
ELB	x		x	
AMA	x			x
TLB	x		x	
NEL	x		x	
SLA		x		x
NWA		x	x	
WCP		x		x
SAP		x		x
WCL		x		x

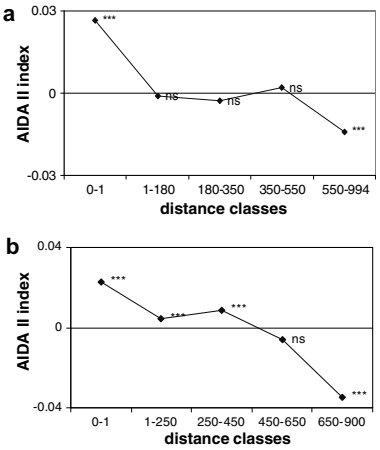


Fig. 4. Spatial autocorrelation analysis as performed by AIDA software: (a) complete Italian dataset; (b) Italian dataset excluding the NWA and CMA samples. $^{***}P < 0.005$; ns, not significant.

2.49% (Φ_{sc} 0.023, $P \ll 0.01$). We then tested a modified grouping (Table 3, PC column) as suggested by PC plot (Fig. 3, positive vs. negative values along axis 1) obtaining a significant percentage of genetic variation between groups of 2.68% (Φ_{ct} 0.026, $P < 0.01$) and within groups 0.88% (Φ_{sc} 0.015, $P \ll 0.01$). The main differences between the two groupings tested (latitude vs. principal component based) are the positioning of the central Italian samples AMA and CMA, and of the southern sample NWA. Starting from the first grouping tested (by latitude), we checked how moving samples across the groups influenced the genetic variation distribution. We moved the samples singularly from one group to the other and estimated the amount of genetic variation between and within groups. The movement of the NWA and CMA samples changed the between group variation to positive

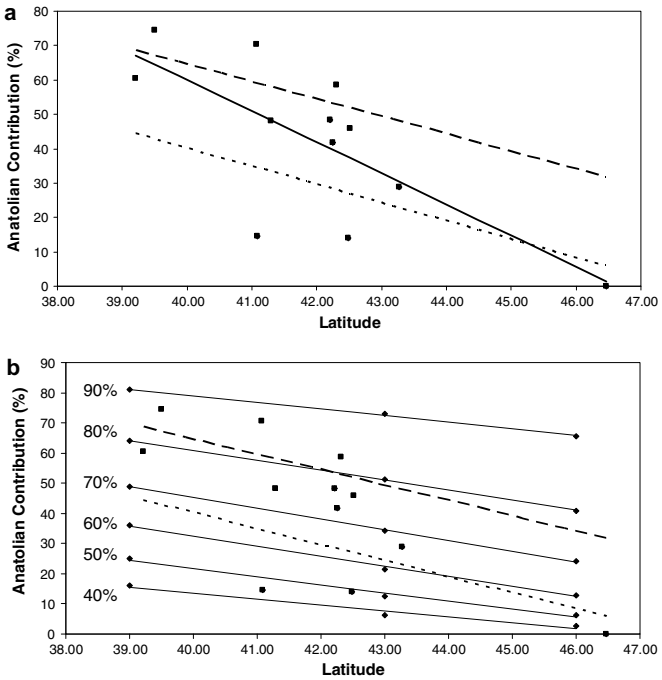


Fig. 5. Admixture analysis: (a) regression lines for the Anatolian genetic contribution to the Italian samples as estimated by the ADMIX software. Dashed line, regression line estimated for southern Italian samples; dotted line, regression line estimated for Northern Italian samples; continuous line, regression line estimated for the entire Italian dataset; (b) regression lines estimated for three admixture event with different newcomer contribution, as described in the Section 3. Admixture proportions are indicated beside each line (see text for description). Dashed and dotted lines are the regression lines estimated for Southern and Northern Italian samples, respectively.

values (0.69% and 1.50%, respectively) while any other shuffling maintained negative values. Moving both at the same time increased the between groups values to 2.89% (Φ_{ct} 0.028, $P \ll 0.01$) and decreased the within groups value to 0.77% (Φ_{sc} 0.007, $P \ll 0.01$). The shuffle of the AMA sample between the two groups did not have any relevant effect. CMA and NWA samples result clearly separated from geographically surrounding samples when investigated by the Barrier 2.2 Software (data not shown).This seems to suggest that for samples east of the Apennines (CMA) the clustering of southern samples is supported for higher latitudes than west of the Apennines. The NWA sample is clearly representing a geographical out-group in respect to the other southern samples (see below). We therefore considered CMA and NWA as part of South and North Italian peninsula respectively in the subsequent analysis. Given the inconclusive indication of the AMOVA analysis for the AMA sample, we decided to follow the PC results and considered this sample as part of the South.

3.3. Spatial correlation analysis

In order to properly test for the presence of clinally distributed genetic variation, we performed spatial correlation analysis as implemented in the AIDA software (Bertorelle and Barbujani, 1995), using hg frequencies. Five classes of distance were selected in order to have similar number of comparisons within each class (Fig. 4a).The results are significant (considering within population diversity) for distances above 500 km. Shorter distance classes were all non significant. We also evaluated the influence of each sample on the observed pattern by excluding the samples one at a time. The exclusion of CMA or the NWA sample resulted in a clinal pattern significant for all distance classes, while the alternative exclusion of the other samples resulted in one or more than one class not being significant (data not shown). The simultaneous exclusion of both CMA and NWA resulted in a very clear clinal pattern (Fig. 4b), confirming the absence of correlation with latitude for these samples (see above).

3.4. Admixture analysis

Given the current proposed models for agricultural technology spread in Europe [CD vs DD; Barbujani and Goldstein, 2004, Currat and Excoffier, 2005], we investigated the genetic impact that Near Eastern populations might have had on Italian populations by admixture analysis performed using the ADMIX program (Bertorelle and Excoffier, 1998). This software estimates the contribution of two source population to the tested population, using allele frequencies. In admixture analysis, a major problem is to correctly identify the source groups. Chikhi et al. (2002) used data of Basques, Lebanese, Syrians and Turks (Semino et al., 2000) to investigate the degree of Near Eastern introgression into a large European dataset by the use of a likelihood approach (Chikhi et al., 2001). Results suggested that all European populations probably experienced Neolithic introgression, with admixture values that must have been larger than 70%. We previously suggested that current Levantine population might have experienced very strong Arab introgression (Capelli et al., 2006b). For this reason, we alternatively selected a large Anatolian data set as representative of newcomer Farmers (Cinnioglu et al., 2004). Linguistic and genetic data support the Basques as descendants from the Palaeolithic inhabitants of Europe (Gamkrelidze and Ivanov, 1990; Cavalli-Sforza et al., 1994; Richards et al., 2000; Semino et al., 2000; Wilson et al., 2001; Chikhi et al., 2002) even if some concerns have been raised (Alonso et al., 2005). However, it has to be noted that drift and bottlenecks might have introduced severe distortions in current Basque populations compared to the ancient ones (Chikhi et al., 2002; Alonso et al., 2005). In order to minimise such occurrences, and following previous results describing limited genetic variation across Iberian samples (Flores et al., 2004), we have decided to pull together data from Iberia (Flores et al., 2004; Bosch et al., 2001) as a representative of the original European inhabitants. The genetic contribution of Anatolian and Iberian source populations were estimated in all the Italian samples by the use of ADMIX. Anatolian contribution estimates were plotted vs. the corresponding sample latitudes. The regression line from these data shows an inverse correlation between Anatolian contribution and latitude (see Fig. 5a). We additionally estimated the regression lines of the samples when clustered according the PC plot results (Table 3, PC column). Interestingly the Northern and Southern samples have two different regression lines, pointing to different degree of genetic impact of the newcomers on the Italian populations. As noted by Chikhi et al. (2002), admixture expectations can be calculated considering a very simple admixture model. This model takes into consideration a number of admixture steps in which the newcomer populations mix with the original inhabitants and create a new community, from which a new "admixture wave" is generated. With P_N the proportion of farmers in the admixed population, the newcomer contribution to each location will decrease in a geometric way from P_N to P_N^n, where n is the number of admixture events (Chikhi et al., 2002). For example, given a P_N of 0.9, and considering a minimum of three admixture events occurring, the proportion of newcomer "genes" will be 81%, 73% and 66%. Following this approach, we plotted the estimated Newcomers contribution vs. three latitude points, evaluating the effect of different P_N admixture proportions (Fig. 5b). If compared to the regression lines estimated from North and South Italian samples, the different position of the two lines appears to be related to different admixture contributions, with Southern samples showing higher contribution than Northern ones (Fig. 5b, Mann–Whitney test, $P < 0.05$).

4. Discussion

Two alternative models have been proposed for the dispersion of agricultural technology in the European continent. The CD–cultural diffusion–model suggests a gradual acceptance of the new technology by the original inhabitants with little or no admixture with Near Eastern farmers. The DD–demographic diffusion–model instead points to substantial genetic introgression of the Near Eastern populations, supported in their demographic expansion by the resources offered by the new food producing technology (Cunliffe, 2001). The two models have clear expectations that can be tested for concordance with the observed data. In the absence of an incoming population, as in the case of the CD model, local drift and gene flow would be the main forces shaping genetic variation. Random fluctuation across populations would prevent the establishment of gradients and no evidence of admixture would be expected. Gene frequency clines have been indicated as necessary but not sufficient to support the DD model (Currat and Excoffier, 2005). The distribution of genetic variation following a population expansion is associated with a loss of genetic diversity due to a succession of small founder effects (Barbujani et al., 1995). The DD model would then predict that along the direction of dispersal, frequency and diversity clines would be generated. Given the population introgression, genetic evidence of admixture is also expected to be found. We extensively sampled along the peninsula to specifically address the issue of the distribution of Y chromosome genetic variation in the light of agriculture diffusion models. When compared to other European samples, no outgroups were found among the Italian samples (Fig. 3). The samples in fact are distributed within the genetic variation shown by other European and Mediterranean populations. However, a limited degree of separation among Italian groups emerged along the first principal component, with Northern Italian samples closer to western European populations and Southern samples closer to South East and South Central European groups, with few exceptions (Fig. 3). This different affinity was also highlighted by AMOVA analysis conducted on microsatellite variation. Similar results were also shown in the seminal work on classical polymorphisms by Piazza et al.

(1988). The Italian samples here analysed appear to be placed within the ES–NW European cline that a number of previous studies have already described and that has been considered as compatible with a demographic scenario of admixture between the Near East farmers and the long-term European Mesolithic inhabitants (Menozzi et al., 1978; Semino et al., 2000; Rosser et al., 2000). We have identified clines for three haplogroups, two of which showed also diversity gradients (Fig. 2).

In the light of Near Eastern gene flow, admixture analysis revealed Anatolian introgression in most of the Italian samples. Considering the expectations related to the different models proposed for spread of agricultural technology, these results support the DD model. Despite the presence of Neolithic genes in the current male Italian population, the admixture values as estimated by ADMIX suggested a differential impact of the newcomers across the Italian samples. The estimated degree of introgression is in fact not consistent across all areas, with Southern samples experiencing higher Anatolian contribution than Northern samples (Fig. 5a). A very rough estimation based on Fig. 5b would suggest a 70–90% contribution for the former and 50–70% for the latter. However, given the number of assumptions, these values should not be taken as absolute. The selection as a second source population of a recently available larger sample of Basques (Alonso et al., 2005) did not change the observed pattern (data not shown). It is interesting to note that only one sample (SAP) displayed an Anatolian genetic contribution not significantly different from 100% (data not shown). Simulated samples obtained by re-sampling either Iberians or Anatolians and tested using the same source populations confirmed the sensitivity of this approach as these simulated samples were not significantly different from 0% and 100% Anatolian contribution, respectively (data not shown). These results underline that, beside geographically different Near Eastern contributions, population replacement was not complete across the peninsula. It follows that both Neolithic and Mesolithic genetic components can be found in current Italian male gene pool.

It is interesting to note that hg R1*(xR1a1) does show a frequency cline, opposite to the ones shown by J2 and E3b1, but apparently no diversity gradient is associated (Fig. 2). We are aware that conclusions drawn on single haplogroups are subject to bias and should not be equated to those drawn from entire samples, but we note that populations do have different hg composition that might retain signatures of past demographic events. The Mesolithic populations had low population density and possibly limited gene flow across groups (Mithen, 2004). If we assume that Mesolithic population were characterised by high frequencies of hg R1*(xR1a1) (as the case in current Basque groups, usually considered as representative of the original inhabitants of Europe) (Semino et al., 2000; Wilson et al., 2001; but see also Alonso et al., 2005), genetic variation within this haplogroup would be independent of geographic sampling and instead mainly shaped by local demographic history. It follows that R1*(xR1a1) diversity would not be expected to show clines related to latitude but instead would be randomly distributed across populations. The later newcomers as represented by Neolithic farmers, would have expanded and admixed with these Mesolithic groups, and generated, as expected, frequency and diversity clines along the direction of dispersal as indeed shown by their most representative chromosome types, E3b1 and J2 (Rosser et al., 2000; Semino et al., 2000, 2004; Cinnioglu et al., 2004). Other less common haplogroups might have retained the signature of those events but the current limited sampling size might have prevented their detection. Along the direction of dispersal (Barbujani et al., 1995), an opposite frequency cline, but not a diversity one, for hg R1*(xR1a1) would be generated. The observed higher Near Eastern contribution to East Apennines vs West Apennines samples for northern latitudes is consistent with the archaeological separation existing among early agricultural areas (Cunliffe, 2001).

The current set of data also provides a first frame for testing the hypothesis of genetic continuity from Palaeolithic to Mesolithic in Italy through the last Ice age. This would point to the presence of an Italian Pleistocene refugium, postulated for Iberian, Italian and Balkan peninsulae for a number of species (Hewitt, 2001; Brito, 2005), but not proposed for humans (Semino et al., 2000; Rosser et al., 2000). This inconsistency could be possibly due to the fact that so far no specific haplogroups have been identified at Y chromosome level in Italy (Semino et al., 2000). However this could be due to lack of resolution in the current set of markers and other Y chromosome sublineages not yet characterised might represent a specific marker for the Italian area. Taking into consideration current European Y chromosome hg distribution and data presented here, a possible candidate could be within hg R1*(xR1a1). A comparison of the genetic variation estimated as the variance of the repeat scores averaged across loci of this group in both Iberia and Italy did not show significant difference ($P > 0.05$, data not shown, Brion et al., 2004; Bosch et al., 2000, 2001). Looking at the pattern of haplotypes sharing within hg P [that contains hg R1*(xR1a1)], only 28% of those are shared among the two populations. This value is well below the one estimated when comparing Iberia with areas re-peopled after last glacial maximum, as the British Isles (47%) (Capelli et al., 2003). The opposite pattern is instead observed when comparing haplotypes within haplogroups whose dispersion was probably associated with different and more recent events, as hg J (data not shown). This suggests that the R1*(xR1a1) variation present in Italy appear not to be a subset of the Iberian one. More extensive analysis would give the opportunity to test this hypothesis further.

Previous studies using autosomal data described the presence of a major North–South Cline within the peninsula. Cavalli-Sforza et al. (1994) showed that 27% of the total genetic variation as shown by classic polymorphisms was summarized along this axis. When compared to European

samples, populations from South Italy clustered with Mediterannean groups, while the others grouped with West and Central European populations (Piazza et al., 1988). Authors suggested the Greek colonization in the South as the major demographic event shaping observed diversity (Piazza et al., 1988) on the basis of compatible historical scenarios. However, this hypothesis was never thoroughly tested, especially in the light of alternative European scenarios proposed by the same authors supporting a Neolithic demic dispersion model (Menozzi et al., 1978; Ammerman and Cavalli-Sforza, 1984; Cavalli-Sforza et al., 1994) and taking in consideration that Greece was the only Mediterranean sample outside Italy included in the PC analysis (Piazza et al., 1988). Assuming the Greek diaspora model, South Italian samples should be genetically closer to Greece than to Anatolia, while the Neolithic model would not show significant differences. We note that D_c and $\Delta\mu$ genetic distances are linearly related to time (Cavalli-Sforza and Edwards, 1967; Goldstein et al., 1995b). We calculated these genetic distances for WCL, WCP and SAP samples vs. Anatolia and Greece (Cinnioglu et al., 2004; Parreira et al., 2002). The three Italian populations were not only closer to Anatolia than to Greece, but all values for Anatolia were smaller than those for Greece (data not shown). This is confirmed also using more specific regional Greek samples (data not shown, Robino et al., 2004). Assuming proper identification of the source populations, these results suggest that in terms of demographic influence on the paternal Italian gene pool, the role of Neolithic farmers was greater than Greek historical colonisers of South Italy.

Similarly, given the sporadic and rare distribution of the E3b2 chromosomes, it is possible to conclude that North African gene flow, if any, left no significant evidence in the current Italian Y chromosome pool (Bosch et al., 2000, 2001; Capelli et al., 2006b).

We finally note that in a recent study, Di Giacomo et al. (2003) genotyped Y chromosome markers for 524 Italians sampled in 17 locations. They found that, excluding R1*(xR1a1), no other clines could be identified and concluded that most of the observed variation was due to drift and founder effects. Local drift has to be expected, due to local demographic histories. This seems the case of the NWA sample, as shown by its reduced genetic variation (Table 2). The NWA sample, despite its localisation in the South, tends clearly to cluster with Northern populations, as shown by various analyses. This is driven by the combination of high hg R1*(xR1a1) frequency and absence/low frequency of hgs E3b1 and J2. In our analyses, some Y chromosome lineages, despite local drift and gene flow, still show the signature of dispersal events in the past. Inspecting the data of Di Giacomo et al., it emerges that despite a lower number of samples points (12 vs. 17) our study is characterised by a larger average sample size, almost twice as much (58 vs. 31). Additionally, while only one of our samples (AMA) is below their average sample size, none of theirs is above or close to our average, with the largest population size in Di Giacomo et al. (2003) being 48. When we included their samples in our analysis, we confirmed the clinal distribution for R1*(xR1a1), J2 and DE hgs (in their study only the YAP marker was genotyped in the DE lineage so this is the closest approximation for E3b1)(data not shown). Besides recognising that drift definitively had a role in shaping current Y chromosome genetic variation, however we concluded that in Italy more than 70% of the observed diversity is distributed along gradients and that Anatolian Farmers did have a different demographic impact on different Italian areas for paternal lienages.

Acknowledgments

We thank all persons and local communities that donated their DNA and made the present study feasible. We additionally thank all the people that were involved in the sampling: Domenico Alfieri, Laura Baldassarri, Fidelia Cascini, Romolo Donnini, Roberto Festa, Leonardo Grimaldi, Armando Mannucci, Sara Partemi, Angela Reveruzzi, Daniela Vantaggiato. We additionally would like to thank the AVIS Blood collection directors of Norma, Giuseppe Santucci and Sezze, Ubaldo Brandolini, Dr. Semeraro, Karl and Elisabeth Mutschlechner, and the City Council of St. Vigil in Ennenberg. C.C. thanks Jim Wilson and Martin Richards for useful comments to an early version of the manuscript. This study was funded by the Italian Ministry of University (PRIN-MIUR 2002, N.2002063871).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2006.11.030.

References

Alonso, S., Flores, C., Cabrera, V., Alonso, A., Martin, P., Albarran, C., Izagirre, N., de la Rua, C., Garcia, O., 2005. The place of the Basques in the European Y-chromosome diversity landscape. Eur. J. Hum. Genet. 13, 1293–1302.

Ammerman, A.J., Cavalli-Sforza, L.L., 1984. The Neolothic transition and the genetics of populations in Europe. Princeton University Press, Princeton.

Barbujani, G., Bertorelle, G., Capitani, G., Scozzari, R., 1995. Geographical structuring in the mtDNA of Italians. Proc. Natl. Acad. Sci. USA 92, 9171–9175.

Barbujani, G., Goldstein, D.B., 2004. Africans and Asians abroad: genetic diversity in Europe. Annu. Rev. Genomics Hum. Genet. 5, 119–150.

Barbujani, G., Sokal, R.R., 1991. Genetic population structure of Italy. II. Physical and cultural barriers to gene flow. Am. J. Hum. Genet. 48, 398–411.

Barbujani, G., Sokal, R.R., Oden, N.L., 1995. Indo-European origins: a computer-simulation test of five hypotheses. Am. J. Phys. Anthropol. 96, 109–132.

Bertorelle, G., Barbujani, G., 1995. Analysis of DNA diversity by spatial autocorrelation. Genetics 140 (2), 811–819.

Bertorelle, G., Excoffier, L., 1998. Inferring admixture proportions from molecular data. Mol. Biol. Evol. 15, 1298–1311.

Bosch, E., Calafell, F., Perez-Lezaun, A., Comas, D., Izaabel, H., Akhayat, O., Sefiani, A., Hariti, G., Dugoujon, J.M., Bertranpetit,

J., 2000. Y chromosome STR haplotypes in four populations from northwest Africa. Int. J. Legal. Med. 114, 36–40.

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

Brion, M., Quintans, B., Zarrabeitia, M., Gonzalez-Neira, A., Salas, A., Lareu, V., Tyler-Smith, C., Carracedo, A., 2004. Micro-geographical differentiation in Northern Iberia revealed by Y-chromosomal DNA analysis. Gene 329, 17–25.

Brito, P.H., 2005. The influence of Pleistocene glacial refugia on tawny owl genetic diversity and phylogeography in western Europe. Mol. Ecol. 14, 3077–3094.

Capelli, C., Wilson, J.F., Richards, M., Stumpf, M.P., Gratrix, F., Oppenheimer, S., Underhill, P., Pascali, V.L., Ko, T.M., Goldstein, D.B., 2001. A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. Am. J. Hum. Genet. 68, 432–443.

Capelli, C., Redhead, N., Abernethy, J.K., Gratrix, F., Wilson, J.F., Moen, T., Hervig, T., Richards, M., Stumpf, M.P., Underhill, P.A., Bradshaw, P., Shaha, A., Thomas, M.G., Bradman, N., Goldstein, D.B., 2003. A Y chromosome census of the British Isles. Curr. Biol. 13, 979–984.

Capelli, C., Arredi, B., Baldassari, L., Boschi, I., Brisighelli, F., Caglia, A., Dobosz, M., Scarnicci, F., Vetrugno, G., Pascali, V.L., 2006a. A 9-loci Y chromosome haplotype in three Italian populations. Forensic Sci. Int. 159, 64–70.

Capelli, C., Redhead, N., Romano, V., Calì', F., Lefranc, G., Delague, V., Megarbane, A., Felice, A.E., Pascali, V.L., Neophytou, P.I., Poulli, Z., Novelletto, A., Malaspina, P., Terrenato, L., Berebbi, A., Fellous, M., Thomas, M.G., Goldstein, D.B., 2006b. Population structure in the Mediterranean basin: a Y chromosome perspective. Ann. Hum. Genet. 70, 207–225.

Cavalli-Sforza, L.L., Edwards, A.W., 1967. Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet. 19 (Supp1. 3), 233.

Cavalli-Sforza, L.L., Menozzi, P., Piazza, A., 1994. The History and Geography of Human Genes. Princeton University Press, Princeton, pp. 277–280.

Cinnioglu, C., King, R., Kivisild, T., Kalfoglu, E., Atasoy, S., Cavalleri, G.L., Lillie, A.S., Roseman, C.C., Lin, A.A., Prince, K., Oefner, P.J., Shen, P., Semino, O., Cavalli-Sforza, L.L., Underhill, P.A., 2004. Excavating Y-chromosome haplotype strata in Anatolia. Hum. Genet. 114, 127–148.

Chikhi, L., Bruford, M.W., Beaumont, M.A., 2001. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics 158, 1347–1362.

Chikhi, L., Nichols, R.A., Barbujani, G., Beaumont, M.A., 2002. Y genetic data support the Neolithic demic diffusion model. Proc. Natl. Acad. Sci. USA 99, 11008–11013.

Cunliffe, B., 2001. The Oxford Illustrated History of PreHistoric Europe. Oxford University Press, Oxford.

Currat, M., Excoffier, L., 2005. The effect of the Neolithic expansion on European molecular diversity. Proc. Biol. Sci. 272, 679–688.

Di Giacomo, F., Luca, F., Anagnou, N., Ciavarella, G., Corbo, R.M., Cresta, M., Cucci, F., Di Stasi, L., Agostiano, V., Giparaki, M., Loutradis, A., Mammi', C., Michalodimitrakis, E.N., Papola, F., Pedicini, G., Plata, E., Terrenato, L., Tofanelli, S., Malaspina, P., Novelletto, A., 2003. Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. Mol. Phylogenet. Evol. 28, 387–395.

Flores, C., Maca-Meyer, N., Perez, J.A., Gonzalez, A.M., Larruga, J.M., Cabrera, V.M., 2003. A predominant European ancestry of paternal lineages from Canary Islanders. Ann. Hum. Genet. 67, 138–152.

Flores, C., Maca-Meyer, N., Gonzalez A, M., Oefner, P.J., Shen, P., Perez, J.A., Rojas, A., Larruga, J.M., Underhill, P.A., 2004. Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: implications for population demography. Eur. J. Hum. Genet. 10, 855–863.

Gamkrelidze, T., Ivanov, V., 1990. The early history of Indo-European. Sci. Am. 262, 110–116.

Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L., Feldman, M.W., 1995a. An evaluation of genetic distances for use with microsatellite loci. Genetics 139, 463–471.

Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L., Feldman, M.W., 1995b. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA 92, 6723–6727.

Goncalves, R., Freitas, A., Branco, M., Rosa, A., Fernandes, A.T., Zhivotovsky, L.A., Underhill, P.A., Kivisild, T., Brehm, A., 2005. Y-chromosome lineages from Portugal, Madeira and Acores record elements of Sephardim and Berber ancestry. Ann. Hum. Genet. 69, 443–454.

Hewitt, G.M., 2001. Speciation, hybrid zones and phylogeography—or seeing genes in space and time. Mol. Ecol. 10, 537–549.

Manni, F., Guerard, E., Heyer, E., 2004. Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. Hum. Biol. 76, 173–190.

Menozzi, P., Piazza, A., Cavalli-Sforza, L., 1978. Synthetic maps of human gene frequencies in Europeans. Science 201, 786–792.

Miller, S.A., Dykes, D.D., Polesky, H.F., 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 16, 1215.

Minch, E., 1996. MICROSAT. Version 1.4. Stanford University Medical Centre, Stanford, USA.

Mithen, S., 2004. After the ice: a global human history 20,000–5000 BC. Phoenix Press, London, UK.

Nasidze, I., Quinque, D., Dupanloup, I., Cordaux, R., Kokshunova, L., Stoneking, M., 2005. Genetic evidence for the Mongolian ancestry of Kalmyks. Am. J. Phys. Anthropol. 128, 846–854.

Onofri, V., Alessandrini, F., Turchi, C., Pesaresi, M., Buscemi, L., Tagliabracci, A., 2006. Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs. Forensic Sci. Int. 157, 23–35.

Parreira, K.S., Lareu, M.V., Sanchez-Diz, P., Skitsa, I., Carracedo, A., 2002. DNA typing of short tandem repeat loci on Y-chromosome of Greek population. Forensic Sci. Int. 126, 261–264.

Pericic, M., Lauc, L.B., Klaric, I.M., Janicijevic, B., Behluli, I., Rudan, P., 2004. Y chromosome haplotypes in Albanian population from Kosovo. Forensic Sci. Int. 146, 61–64.

Piazza, A., Cappello, N., Olivetti, E., Rendine, S., 1988. A genetic history of Italy. Ann. Hum. Genet. 52, 203–213.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Golge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A., Papiha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., Di Rienzo, A., Novelletto, A., Oppenheim, A., Norby, S., Al-Zaheri, N., Santachiara-Benerecetti, S., Scozari, R., Torroni, A., Bandelt, H.J., 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. Am. J. Hum. Genet. 67, 1251–1276.

Robino, C., Gino, S., Ricci, U., Grignani, P., Previdere, C., Torre, C., 2002. Y-chromosomal STR haplotypes in an Albanian population sample. Forensic Sci. Int. 129, 128–130.

Robino, C., Varacalli, S., Gino, S., Chatzikyriakidou, A., Kouvatsi, A., Triantaphyllidis, C., Di Gaetano, C., Crobu, F., Matullo, G., Piazza, A., Torre, C., 2004. Y-chromosomal STR haplotypes in a population sample from continental Greece, and the islands of Crete and Chios. Forensic Sci. Int. 145, 61–64.

Roewer, L., Croucher, P.J., Willuweit, S., Lu, T.T., Kayser, M., Lessig, R., de Knijff, P., Jobling, M.A., Tyler-Smith, C., Krawczak, M., 2005. Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Hum. Genet. 116, 279–291.

Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., Beckman, G., Beckman, L., Bertranpetit, J., Bosch, E., Bradley, D.G.,

Brede, G., Cooper, G., Corte-Real, H.B., de Knijff, P., Decorte, R., Dubrova, Y.E., Evgrafov, O., Gilissen, A., Glisic, S., Golge, M., Hill, E.W., Jeziorowska, A., Kalaydjieva, L., Kayser, M., Kivisild, T., Kravchenko, S.A., Krumina, A., Kucinskas, V., Lavinha, J., Livshits, L.A., Malaspina, P., Maria, S., McElreavey, K., Meitinger, T.A., Mikelsaar, A.V., Mitchell, R.J., Nafa, K., Nicholson, J., Norby, S., Pandya, A., Parik, J., Patsalis, P.C., Pereira, L., Peterlin, B., Pielberg, G., Prata, M.J., Previdere, C., Roewer, L., Rootsi, S., Rubinsztein, D.C., Saillard, J., Santos, F.R., Stefanescu, G., Sykes, B.C., Tolun, A., Villems, R., Tyler-Smith, C., Jobling, M.A., 2003. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am. J. Hum. Genet. 67, 1526–1543.

Schneider, S., Roessli, D., Excoffier, L., 2000. ARLEQUIN ver. 2.000: A Software for Population Genetic Data Analysis. Genetics and Biometery Laboratory, University of Geneva, Geneva.

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

Semino, O., Magri, C., Benuzzi, G., Lin, A.A., Al-Zahery, N., Battaglia, V., Maccioni, L., Triantaphyllidis, C., Shen, P., Oefner, P.J., Zhivotovsky, L.A., King, R., Torroni, A., Cavalli-Sforza, L.L., Underhill, P.A., Santachiara-Benerecetti, A.S., 2004. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am. J. Hum. Genet. 74, 1023–1034.

Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J., Barbujani, G., 2000. Geographic patterns of mtDNA diversity in Europe. Am. J. Hum. Genet. 66, 262–278.

Slatkin, M., A measure of population subdivision based on microsatellite allele, frequencies, 1995. Genetics 139, 457–462.

Thomas, M.G., Bradman, N., Flinn, H.M., 1999. High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Hum. Genet. 105, 577–581.

Y Chromosome Consortium., 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. 12, 339–348.

Wilson, J.F., Weiss, D.A., Richards, M., Thomas, M.G., Bradman, N., Goldstein, D.B., 2001. Genetic evidence for different male and female roles during cultural transitions in the British Isles. Proc. Natl. Acad. Sci. USA 98, 5078–5083.

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R.S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., Li, P., Yuldasheva, N., Ruzibakiev, R., Xu, J., Shu, Q., Du, R., Yang, H., Hurles, M.E., Robinson, E., Gerelsaikhan, T., Dashnyam, B., Mehdi, S.Q., Tyler-Smith, C., 2003. The genetic legacy of the Mongols. Am. J. Hum. Genet. 72, 717–721.

PLoS ONE

Research article; June 2012

Population structure of modern-day Italy as revealed by a comprehensive analysis of uniparental genetic variation

Francesca Brisighelli[1,2,3,#], Vanesa Álvarez-Iglesias[1], Manuel Fondevila[1], Alejandro Blanco-Verea[1], Ángel Carracedo[1,4], Vincenzo L Pascali[2], Cristian Capelli[3], Antonio Salas*[1#]

[1]Unidade de Xenética, Facultade de Medicina, Instituto de Medicina Legal, Universidade de Santiago de Compostela, Galicia, Spain; [2]Forensic Genetics Laboratory, Institute of Legal Medicine, Universita` Cattolica del Sacro Cuore, Rome, Italy; [3]Department of Zoology, University of Oxford, Oxford, UK; [4] Fundación Pública Galega de Medicina Xenómica (FPGMX-SERGAS), CIBER enfermedades raras, Santiago de Compostela, Galicia, Spain

Keywords: Italy, Etruscans, Toscana, African slave trade, haplotype, haplogroups, SNPs

#Both authors contributed equally to this work

***Correspondence:** Antonio Salas; Unidade de Xenética, Departamento de Anatomía Patológica e Ciencias Forenses, Universidade de Santiago de Compostela, Galicia, Spain. Tel: +34–981–582327; Fax: +34–981–580336; E-mail: antonio.salas@usc.es

Abstract

Background: Historical documentation reveals the Italy has been a melting point of different Mediterranean cultures and populations. Although many genetic studies have been undertaken in Italy, genetic patterns have never been analyzed comprehensively, including uniparental and autosomal markers throughout the country.

Methods/Principal findings: A total of 583 individuals were sampled from across the Italian Peninsula, representing 12 different populations, two (Ladins, Grecani Salentini) being linguistic isolates. All samples were analyzed for the mitochondrial DNA (mtDNA) control region and selected coding region SNPs (mtSNPs). This data was pooled for analysis with 3,778 mtDNA control-region profiles collected from the literature. A set of Y-chromosome SNPs and STRs were also analyzed in 479 individuals together with a panel of autosomal ancestry informative markers (AIMs) from 441 samples. While some genetic differentiation exists along the Italian Peninsula, the Ladins showed the most distinctive phylogeographic patterns. Data suggest that clinal latitudinal patterns along continental Italy could have been generated by demographical movements occurring during the Neolithic. The Neolithic contribution was estimated for the Y-chromosome as 14.5% and for mtDNA as 10.5%. Y-chromosome data showed larger differentiation between North, Center and South than mtDNA. AIMs detected a minor sub-Saharan component; this is however higher than for other European non-Mediterranean populations. This sub-Saharan contribution was also detected in uniparental markers. Bayesian-based admixture analysis of mtDNA data showed a 27% North African contribution, while the Middle East contributed about 28%. However, these estimates warrant further confirmation, provided the figures are inflated due to the limited molecular resolution provided by available mtDNA control region data.

Conclusions/Significance: Italy shows patterns of molecular variation mirroring other European countries, although some heterogeneity exists based on different analysis and molecular markers. From North to South, Italy shows clinal patterns that were most likely modulated during Neolithic times.

Introduction

Italy has historically been a convenient destination for human populations migrating from Africa, the Middle East and European locations, in part due to the geomorphological characteristics of the Italian Peninsula [1]. These groups settled preferentially on the islands and coastal territories [1]. During the Paleolithic, the icecap expansion of the Late Glacial Maximum (LGM) pushed southward into Italy groups of hunters living in Central European areas [1]. During the Copper, Bronze and Iron ages, few migrations and exchanges occurred between the Mediterranean basin and the Near East [2]. Exchange of metals would determine the transformation of the first social organizations in ancient civilizations [2]. Sardinia, Sicily and Tuscany were among the first Italian territories to be occupied by humans due to their strategic location and the presence in their territories of important metal resources [3].

Different cultures, recognized on the basis of different archeological findings, settlements and burial traditions, arose in the period between the Mesolithic and Iron Age. Before the Roman conquest, ancient Italy was characterized only by the presence of Indo-European populations [4] living in the Italian Peninsula since the second millennium BC, corresponding to the period between the Iron Age and Romanization. All these populations are generally known as Italics.

The record of all the populations that inhabited the Italian territory during (pre)-history is incomplete; many records were of uncertain location and/or ambiguous denomination [4]. At the time of the Roman Empire, at least two non-Indo-European populations still inhabited Italy, namely, the Ligures, in the northwestern area (between the rivers Arno and Rhone; in a wider area than present-day Liguria), and the Etruscans with settlements located in areas far from the Etruria (Tuscany and High Latium), such as the Po Plain and the coast of Campania. Throughout the sixth century BC, Etruscans represented the community in Italy with the most advanced organization. At the same time, Sardinia experienced the flourishing of a non-Indo-European Nuragic civilization and, then, the Phoenician colonization

Genetics alone cannot disentangle the extremely complex demography of Italy through history. Some demographic movements have however left signals on uniparental and nuclear markers. Most of the genetic studies targeted local, e.g. [5], or regional, e.g. [6-9], Italian populations. For the Y-chromosome, some attempts have been undertaken to

analyze Italian variation to a more general scale [10-12]. Many studies have analyzed specific haplogroups in the Y-chromosomes, e.g. [13,14], or the mtDNA, e.g. [6,7]. In general, the different studies indicate that the genetic structure of the present Italian population seems to reflect, at least in part, the ethnic stratification of pre-Roman times [12]. Studies carried out in the past appear to show a major North–South cline consistent with archaeological estimates of two distinct processes: the first colonization of the area during the Paleolithic period and the subsequent Neolithic expansion from the Middle East after the last glacial [12]. There is some correspondence between patterns of variation at the Y-chromosome and geography. Thus, northern Italy shows similar frequencies as the haplogroups of Central Europe, with prevalence of the western R1-M173 haplogroup in respect to the eastern I-M170. In the North, E3b1-M35 and J2-M172 show low frequencies but are more prevalent in the South, which has been interpreted to be a signal of the gene flow coming from Central European Neolithic farmers [15]. R1a1-M17 is rather rare, both in the North, where it probably originates from eastern Europe, and in the South, of possible Greek provenience [15]. Occurrence of J2-M172 Y-chromosomes in Tuscany has been related to the Etruscan heritage of the region (see [15]). The two Italian major islands, Sicily and Sardinia, show a different demographic history. The Y-chromosome variability of Sicily shares a common history with that of southern Italy, enriched by an additional Arab contribution, but also North African and Greek influences [16]. On the other hand, Sardinia has been considered to be a genetic outlier within Europe showing clear signals of founder effects; some scholars suggest that its peoples could be of ancient Iberian origin [17]; recent genetic studies point to genetic contribution coming from southern France [18]. Mitochondrial DNA studies show that Italy does not differ too much from other European populations; however, some populations have the same peculiarities and preserve signals of the ancient past demographic event, such as the Tuscans [6,7], or the Ladins [5,19,20]. Recently, patterns of variation observed in haplogroup U5b3 demonstrated for the first time the existence of a North Italian pre-historical human refuge from the hostile Central European regions covered by the ice of the Last Glacial Maximum period [18]; this area, as it was also the Franco-Cantabrian region [21-24], served as a region of European repopulation during the beginning of the Holocene.

The main aim of the present study was comprehensively to analyze the patterns of mtDNA and Y-chromosome variation in Italy. This study differs from previous ones in that: (1) it provides mtDNA data from 12 new sample populations from Italy; (2) we analyzed two linguistic isolates, Ladin and Grecani Salentini, the latter sampled for the first time in this study; (3) we analyzed a sample population from Lucera (Southern Italy) for the first time, a population that according to documentation received an important input of North African immigrants during the thirteenth century; (4) it analyzed the patterns of mtDNA variation in Italy globally, that is, by combining more than 3,700 control region profiles from the literature (41 population samples in total) coupled with the more than 580 new profiles provided here; (5) Y-chromosome haplotype and haplogroup patterns are analyzed in parallel with the mtDNA data in order to determine the possible differences that occurred historically in the male versus female demographic movements; and (6) the influx of migrants from Africa (North and sub-Saharan) and other regions is also analyzed using phylogeographic inferences, and also a model of admixture based on haplotypic data and a panel of ancestry informative markers (AIMs).

Material and methods

Ethics statement

Written informed consent was obtained from all sample donors. Analysis of mtDNA sequences was approved by the institutional review boards of the Università Cattolica del Sacro Cuore (Roma). Moreover, the study conforms to the Spanish Law for Biomedical Research (Law 14/2007- 3 of July).

Samples

A total of 583 individuals were sampled from along the Italian Peninsula, representing 12 different populations (**Figure 1**), two of them (Ladin and Grecani Salentini) being linguistic isolates, and the Lucera being a historical enclave of Arabs coming from North Africa. A brief description of these latter three populations is given below.

In the Italian territory, the Alpine arc represents one of the main areas of presence of alloglot populations, some of them biologically isolated for historical and geographic reasons [25]. At the end of the medieval period (~1200 AD) and especially in the valley zone, a first colonization of native peasants began, starting with the use of lands previously exploited

only for pasture and the lumber. Successively, with different modalities and under the control of laic and ecclesiastical owners, the colonization process involved migrant nuclei from the Tyrol, Carinthian area and other zones. These migrants first filled uncultivated spaces, and then moved away, creating new settlements forming "ethnic islands", above all those of the germanophone culture, which nowadays still exists [26]. Currently, the alpine arc populations are differentiated with a remarkable cultural diversity that is well represented by linguistic elements. In this territory, besides the official main languages, numerous minority languages or dialects are also the cultural patrimony of linguistic minorities [25,27]. Ladin is often attributed to be a relic of vulgar Latin dialects associated with Rhaeto-Romance languages. Starting in the sixth century, the Bavarii migrated in from the North, while from the South the Italian language started to push northwards, which further shrank the original extent of the Ladin area. Only in the more remote mountain valleys was Ladin able to survive. In the vast multi-ethnic Holy Roman Empire, and then after 1804 the Austrian empire, the Ladins were left in relative peace and were allowed to continue the use of their language and culture.

Grecani Salentini is a Hellenic-speaking linguistic island of Salento, situated in southern Puglia, and consisting of nine municipalities in which a neo-Greek dialect, also known as Grecanic or Griko, is spoken. The origins of this linguistic island in Salentine Greece are uncertain. The German linguist G. Rohlfs proposed its origin in the Magna Graecia region; while O. Parlangeli suggests a byzantine derivation of the Griki of Salento. Greek researchers (e.g. A. Karanastasis) claim the input of byzantine elements in the pre-existing Magna Graecia matrix. The Greek arrival in the Salentine Peninsula occurred both in the ancient age (Magna Graecia), and posterior byzantine dominations. The numerous villages of Grecani Salentini had a Greek culture and language and practiced the Greek-orthodox religion. In the beginning of the Norman conquest (eleventh century), and more intensively with the arrival of different casati (clans) (Svevian, Angioin, Aragones, etc), the catholic clergy and supplanted those of the orthodox faith [28].

The Lucera population has received an important influx from North African Arab peoples (see [29]). Thus, after the collapse of the Roman Empire in Europe, the Arab domination spread into the Mediterranean Basin. Referred to either as Moors in Iberia or Saracens in Southern Italy and Sicily, Arabs arrived in Europe in 711 AD, and in 831 AD Iberia and Sicily were almost completely subjected to Arab domination [29]. In the

thirteenth century, Frederick II moved the Sicilian Arabs to the city of Lucera (North Apulia) [30]. This sample was genotyped for STRs and Y-chromosome SNPs in Capelli et al. [29]

To the best of our knowledge, all individuals collected in the present study were not maternally and paternally closely related; they had different surnames and all the donors referred back at least two generations in the region where the samples were collected.

All the samples were analyzed for the control region and selected mtSNPs (see below). A subset of the samples comprised unrelated males (n = 292) representing seven different populations. These samples were genotyped for a panel of 17 Y-chromosome SNPs (see below), and were previously genotyped for the Yfiler [31]. In addition, autosomal ancestry informative markers (AIMs) were genotyped in 441 individuals (see below).

DNA extraction

Blood extraction was performed with a salting-out method [32], modified and re-adapted to buccal cells. Swabs were incubated in 500 µl of 0.2 sodium acetate, 35 µl of 10% SDS and 20 µl of 20mg/ml Proteinase K for 16 hours at 56 °C. They were then removed and 500 µl of 3 M NaCl solution was added. Proteins were removed by centrifugation, and the DNA precipitated by adding 1 ml of ethanol 100% at −20 °C for a few hours. After centrifugation, the DNA pellet was twice washed with ethanol 70%, dried and re-suspended in water. For the blood samples, aliquots of 500 µl each were thawed and red cells selectively lysed by a 1 x lysis buffer. After three washes with the lysis buffer, white cells were pelleted and the DNA extracted using the salting-out protocol. All the samples were quantified by direct comparison with standard on agarose 1% minigels (1 g of agarose in 100 ml of TBE 1X- from the 1:10 diluition of TBE 10X).

PCR and mtDNA control region sequencing

MtDNA has been sequenced for the complete control region, from position 16024 (in HVS-I) to 569 (in HVS-II). The first and second hypervariable regions (HVS-I/II) were amplified via the polymerase chain reaction (PCR) and using primers reported by Álvarez-Iglesias et al. [33].

PCR was carried out in a 25 µl reaction mix with 1 x reaction buffer (20 mM Tris-HCl, ph 8.0, 0.1 mM EDTA, 1 mM DDT, 50% (v/v) glycerol), 1.5 mM MgCl2, 200 mM

each dNTP, 0.4 μM each primer, 2.5 U (Units). Taq polymerase and 0.1-1 ng DNA template was added to the reaction mixture (Taq DNA Polymerase, recombinant. INVITROGEN® Corporation). Amplification was carried out in a GENE AMP® PCR SYSTEM 9700 (Applied Biosystems, Foster City, California,U.S.A.) using a hot start at 95 °C for 1 min, followed by 36 cycles at 95 °C for 30 sec, 55 °C for 60 sec, and 72 °C for 30 sec and a final extension at 72 °C for 15 min. Before the sequencing reaction, PCR products were checked by electrophoresis in polyacrylamide non-denaturing gel (T9, C5), and subsequently the gel was stained with silver nitrate. PCR products were then purified with a MultiScreen® PCRμ96 Plate (Millipore, Bedford, Ma 01730, U.S.A), 96-well device.The vacuum-based, size exclusion separation effectively and quickly removed the containing salts, unincorporated dNTPs and primers from PCR reactions. Cycle sequencing was performed on both strands in a GENE AMP® PCR SYSTEM 9700 (AB) thermal cycler using the ABI Prism® dRhodamine Terminator Cycle Sequencing Ready Reaction Kit (AB). This kit consists of a reaction mix composed of: DNA-modified and termostable polymerase, Buffer Tris-HCl (pH 9.0), MgCl2, dNTPs, dichlororhodamine-marked ddNTPs. An aliquote of 30 ng amplicon and 3.2 ρM primers were added to a 2 μl reaction mix. Sequencing was carried out using a hot start at 96 °C for 4 min, followed by 36 cycles at 96 °C for 15 sec, 50 °C for 10 sec, 60 °C for 2 sec and a final extension at 60 °C for 10 min. The removal of excess dideoxy terminators, primers and buffer was accomplished with an alcoholic purification.

The sequence products were denatured with deionized formamide and analyzed by capillary electrophoresis on an ABI PRISM 3130® Genetic Analyzer (AB).The resulting data were analyzed with PE/ABD software Sequencing Analysis 5.2 and sequences were aligned and compared with the Cambridge sequence [34] from position 16024 to16569 for HVS-I and from position 1 to 600 for HVS-II by the SeqScape v.2.0 (AB).

Analysis of mtDNA coding region SNPs

Biallelic markers were genotyped using a multiplex approach [35]. The selected SNPs were combined into two multiplex reactions. Multiplex 1 included a selection of SNPs defining common European haplogroups [36]. Multiplex 2 included exclusively polymorphisms defining sub-lineages inside haplogroup H. Primers were designed in order to adjust the annealing temperatures and amplicon lengths to allow analysis in multiplex reactions [35]. The sizes of the PCR products ranged from 80 to 224 bp.

Both multiplexes were performed using 10 ng of DNA template in a 25 µl reaction volume comprising 1 x Taq Gold Buffer (AB), 200 µM of each dNTP, 2 mM MgCl2 and 0.5 U of AmpliTaq Gold Polymerase (AB). For the primer concentrations, see [35].

Amplification was carried out using a GENE AMP® PCR SYSTEM 9700 (AB) thermocycler. After a 95 °C pre-incubation step for 11 min, PCR was performed for a total of 32 cycles using the following conditions: 94 °C denaturation for 30 sec, annealing at 60 °C for 30 sec and extension at 72 °C for 1 min, followed by a 15 min final extension at 72 °C. PCR products were checked by polyacrylamide gel electrophoresis (T9, C5) visualized by silver staining.

After amplification, PCR products required purification to remove primers and unincorporated dNTPs. Post-PCR purification was performed with ExoSapIT (Amershan Pharmacia Biotech): 1 µl of PCR product was incubated with 0.5 µl of ExoSapIT for 15 min at 37 °C followed by 15 min at 80 °C for enzyme inactivation. The minisequencing reaction was performed in a GENE AMP® PCR SYSTEM 9700 (AB) thermocycler following the recommendations of the manufacturer: 2 µl of SNaPshot ready reaction mix, 0.2 µM of extension primer for each SNP (see [35]) and 1 µl of both purified PCR products in a total volume of 7 µl. The reaction mixture was subjected to 25 single base extension cycles of denaturation at 96 °C for 10 sec, annealing at 50 °C for 5 sec and with an extension at 60 °C during 30 sec. After minisequencing reactions, a post-extension treatment to remove the 5'-phosphoryl group of ddNTPs aided the prevention of co-migration of unincorporated ddNTPs with extended primers and production of a high background signal. The final volume (7 µl) was treated with 0.7 µl of SAP (Amersham Biosciences) for 60 min at 37 °C, followed by 15 min at 80 °C for enzyme inactivation.

The minisequencing products (1.5 µl) were mixed with 10 µl of HiDi™ formamide and 0.2 µl of GeneScan-120 LIZ size standard (AB) and electroforesis was performed on an ABI PRISM 3130® Genetic Analyser (AB). The resulting data was analyzed with Gene Mapper ID.

Minisequencing of SNPs characterizing additional typical European haplogroups

Samples that were determined (using the SNP panel above) as being derived from J/T (T14766C; C7028T; T4216C), U (T14766C; C7028T; A12308G) and the U-subclade K (T14766C; C7028T; A12308G; A10398G), were further genotyped using an additional set of

14 haplogroup-specific SNP markers that identify the following sub-branches: J1 (G3010A), J1b (G3010A; C13879T), J1c (G3010A; C114798T), J2 (G15257A), T2a (A14687G), T2b (G5147A), U5a (A14793G), U5a1 (A14793G; A15218G), U5b (A7768G), U5b1 (A7768G; A5656G), U5b2 (A7768G; C1721T), K1 (T14798C; T1189C), K1a (T14798C; T1189C; C0497T) and K2 (T14798C; T1189C; T9716C). PCR and minisequencing reactions were performed as described above. For PCR and minisequencing primer concentrations, see **Table S1**.

Genotyping of Y-SNPs

Biallelic markers were genotyped using a multiplex approach [37]. A set of 30 SNPs was tested, allowing assignation of the analyzed Y-chromosome to haplogroups (Hg), following the nomenclature and the phylogenetic relationships defined from the Y Chromosome Consortium [38]. The selected method for allele discrimination was a single base extension reaction using the SNaPshot multiplex kit (AB). We added the M269 marker to the first of the four multiplexes, in order better to dissect the sub-haplogroup R1b (R1b3). The primers of this marker were M269-F 5'-TCA TGC CTA GCC TCA TTC CT-3' and M269-R 5'-TCT TTT GTG TGC CTT CTG AGG-3', and the minisequencing primer 5'-GGA ATG ATC AGG GTT TGG TTA AT-3'.

Genotyping of AIMs

A panel of 52 AIMs were genotyped according to Sánchez et al. [39] in a subset of 441 individuals. Several other population datasets were used for inter-population comparisons. This data corresponded to the CEPH panel (http://www.cephb.fr/en/cephdb/) as reported in HapMap (http://hapmap.ncbi.nlm.nih.gov/) and was collected using the data-mining tool SPSmart [40,41]; it includes population samples from all over the world (Africa, Europe, Asia, etc.); see legend of **Figure 2** for more information.

Statistical analysis

A total of 42 Italian population samples were analyzed for mtDNA in the present study. Comparative inter-population analyses were also carried out for the HVS-I segment ranging from 16024 to 16365, since this is the analyzed segment common to all of them. Haplotype (H) and nucleotide diversity (π) and other diversity indices [42-44] were

computed using DnaSP 4.10.3 software [45]. Problematic variation located around 16189, usually associated to length heteroplasmy e.g. 16182C or 16183C, was ignored. Analysis of molecular variance (AMOVA) was carried out using Arlequin 3.5. [46]. Nomenclature of mtDNA lineages followed previous studies e.g. [21,23,36,47,48]; see Phylotree for a compilation of the worldwide phylogeny and an update of the nomenclature based on entire mtDNA genomes [49]. Genotyping and documentation errors were monitored following the phylpogenetic principles previously applied e.g. [50-57].

Mitochondrial DNA and Y-chromosome data was collected from the literature. The mtDNA data generated in the present study was analyzed together with 3,834 mtDNA HVS-I Italian profiles collected from the literature (**Table S2**; 76 sample populations). In addition, for some analyses, we collected a database of mtDNA HVS-I profiles, including 23,629 from the profiles of non-Italian Europe (representing 50 population samples), 4,805 from the Middle East (34 population samples), 3,830 from North Africans (10 population samples), and 9,866 Sub-Saharans (40 population samples). The Y-SNPs were analyzed together with 1,251 Italian profiles reported in the literature (16 population samples). A full list of references for all the data used in the present study is given in **Table S2**.

Haplogroup frequencies were estimated by chromosome counting. Statistical differences in haplogroup frequencies were evaluated using a Pearson's chi- square test and by setting up the nominal significant value α as 0.05.

Finally, classification of mtDNA sequences into haplogroups was performed following phylogenetic criteria (Phylotree Build 14, http://www.phylotree.org/) and using both the control region sequence profile and mtSNPs.

Results

Molecular diversity of mtDNA and Y-chromosome Italian profiles

Diversity indices were computed for all the populations analyzed in the present study and also in those Italian populations samples reported in the literature (**Tables 1** and **2**). Population samples were also grouped in main regions (North, Central, South, West, and East) in order to investigate the role of geography in the distribution of mtDNA variation.

Mitochondrial DNA haplotypes for the samples analyzed in the present study are reported in **Table S3**. **Table 1** shows the molecular diversity values based on mtDNA data for 41 Italian population samples. The values indicate that the Isle of Elba is, by far, the

Italian population sample that shows the lowest diversity for all the indices computed, probably as a consequence of its relative isolation from the country. It has been reported that this was a well-known enclave of Etruscan influence, and some mtDNA particularities have been described before [6,7]. Excluding the Isle of Elba, haplotype diversity in Italy ranges from 0.834 to 1, nucleotide diversity from 0.01003 to 0.02409, and the average value of nucleotide differences from 3.4 to 8.19 (a value that is correlated with the nucleotide diversity). In general, Italy shows some level of heterogeneity when examined for diversity values.

When grouping populations by main geographical regions, it can be observed that Central Italy has slightly lower values than North and South Italy for all the indices computed (**Table 1**). The higher diversity values were found in South Italy. Diversity values are however very similar when examining populations located in West Italy versus those in the East. The inclusion of Sicily (as part of South Italy) in the computation does not substantially change these estimates (**Table 1**).

Y-SNP data were obtained for all the samples analyzed in the present study (**Table S4**). **Table 2** shows the diversity indices for the Y-SNPs in different Italian populations. The Y-STR diversity values for the samples analyzed in the present study and other Italian and European samples have already been reported in Brisighelli et al. [31]. As expected, diversity values of Y-SNP haplogroup patterns are lower than those obtained for the mtDNA haplotypes given that the indices are based on haplogroup and not on Y-STR haplotypes. In fact, values based on Y-STR profiles (minimum or extended Yfiler profiles) [31] are higher than those observed for the HVS-I profiles. Ladins are among the populations with the lowest Y-SNP diversity values, while the Grecani Salentini show diversity values that are comparable to other Italian samples. Modena shows notable low haplotype diversity values.

Phylogeography

The mtDNA haplogroup make-up of Italy as observed in our samples fits well with expectations in a typical European population. Thus, most of the Italian mtDNAs (~89%) could be attributed to European haplogroups H (~40%), I (~3%), J (~9%), T (~11%), U (~20%; U minus U6), V (~3%), X (~2%) and W (~1%); Figure 1. There are however important differences in haplogroup frequencies when examining them by main geographical

regions. Thus, for instance, haplogroup H is 59% in the North, 46% in the Center, and decays to ~33% in the South; moreover, these regional differences are statistically significant: North vs South (Pearson's chi-square, unadjusted-P value < 0.00003), and Center vs South (Pearson's chi-square, unadjusted-P value < 0.03724).

Mitochondrial DNA haplotypes of African origin are mainly represented by haplogroups M1 (0.3%), U6 (0.8%) and L (1.2%); from here onwards, L will be used to refer to all mtDNA lineages, excluding the non-African branches N and M [58,59].

A total of 282 Y-chromosomes were analyzed for a set of Y-SNPs and were classified into 22 different haplogroups (**Figure 3**). Two haplogroups were not found, even if markers defining these clades were tested: N3 and R1a1. Five haplogroups represented 76.71% of the total chromosomes: R1b3, J2, I(xI1b2), E3b1 and G. The frequencies averaged across populations were 26%, 21.2%, 10.2%, 9.9% and 9.2%, respectively. The remaining haplogroups sum to 23.2% in the total sample, and never above 4% in single population samples.

R1b3 frequency was found to be higher in the northern part of the country, while G and E3b1, J2 and K2 frequencies were higher in the south and in the central part of the country, respectively (**Figure 1**).

Regional differences are substantially higher in the Y-chromosome than in the mtDNA. Thus, for instance, haplogroup R in the Y-chromosome was 54% in the North, 18% in the Center, and 31% in the South. Frequency differences were statistically significant between North vs Center (Pearson's chi-square, unadjusted-P value = 0.0014), and North vs South (Pearson's chi-square, unadjusted-P value < 0.00004). Haplogroup J2 also revealed important regional differences; it added to 9% in the North, 37% in the Center, and 22% in the South, with statistically significant differences between the North vs Center (Pearson's chi-square, unadjusted-P value < 0.00002), North vs South (Pearson's chi-square, unadjusted-P value < 0.00148), and in the limit of significance Center vs South (Pearson's chi-square, unadjusted-P value < 0.049).

Autosomal ancestry in Italy

A panel of 52 AIMs was genotyped in 435 Italian individuals in order to estimate the proportion of ancestry from a three-way differentiation: sub-Saharan Africa, Europe and Asia. Structure analyses allowed us to infer membership proportions in population samples,

and these proportions can be graphically displayed, as in **Figure 2**. This analysis indicated that Italians have a basal proportion of sub-Saharan ancestry that is higher (9.2%, on average) than other central or northern European populations (1.5%, on average). The amount of African ancestry in Italians is however more comparable to (but slightly higher than) the average in other Mediterranean countries (7.1%). Figure 2 shows in a triangle plot the relationships of Italians compared to other European, African and Asian populations.

PCA observations confirmed the results from Structure analysis, clustering Italian profiles tightly with other European ones. Thus, PCA indicated that North, Central and South Italy do not show differences between them, nor from other European populations (**Figure 2**). PCA also indicated clear-cut differences between Italians, Africans and Asians (**Figure 2**).

Admixture proportions could be computed only on the mtDNA, given the availability of data for other continental locations. We used HVS-I data and followed the method in [60,61]; the following source populations were considered: Europe, the Near East, North East and North West Africa, and sub-Saharan Africa. As expected, the greater proportion was attributed to Europe (P0 = 36%), followed by East and West Africa and the Middle East in similar proportions (P0 ≈ 20%) (**Table 3**). If we pooled East and West Africa in a single North African population, the contribution from Europe increased to 44%, while Near East and North Africa each contributed about 27%. The contribution of sub-Saharan Africa was only about 2%, in agreement with estimates of membership obtained using Structure analysis. These estimates warrant further confirmation, provided the figures are inflated due to the limited molecular resolution provided by available mtDNA control region data.

AMOVA

AMOVA analyses were carried out following different grouping schemes. The samples were pooled into a single population, but also by considering main Italian regions. Analyses were carried out over haplogroups and haplotypes of the Y-chromosome and the mtDNA (**Table 4**).

AMOVA indicated that, among populations, variance was more strongly stratified for the Y-chromosome than for the mtDNA; the difference was much more marked for the analysis based on haplogroups (14.39% vs 1.17%) than for the analysis based on haplotypes (2.34% vs 0.79%). Among population variance was very low when analyzing main

geographical regions; however, it was the latitude (North vs Center vs South) that appeared to account for higher values of among-population variance rather than longitude (West vs East), with the exception of the Y-chromosome haplogroups (although the values are below 1%); **Table 4**. Again, the Y-chromosome showed slightly higher values of among-population variance than did the mtDNA. For the Y-chromosome, a significant proportion of the within-population variance moved to among-population within-groups variance, probably due to the fact that all population samples had a very high proportion of singleton Yfiler haplotypes, elevating the maximum values of haplogroup diversity for all of them [31].

Linguistic isolates: Ladin and Grecani Salentini

Two linguistic isolates are represented in the samples analyzed in the present study: the Ladin and the Grecani Salentini.

Other population samples of the Ladin have already been analyzed in the literature [20,62,63]. We here sampled 41 new individuals from the locality of Val Badia. As reported in **Table 5** for the mtDNA, Val Badia Ladins showed relatively high nucleotide diversity patterns compared to other Ladin populations, but intermediate haplotype diversity values. Compared to other Italian populations, diversity in Ladin populations is generally lower (**Table 1**). For Y-chromosome haplogroups, the differences between Ladin and the rest of Italy were more evident, with the Ladin showing much lower values than average Italians.

The differences between Ladin and other populations were more evident when examining haplogroup frequency patterns (**Figure 4**). The frequency of haplogroup H (58%) was above the frequency of H in North Italy (55%), and was extremely high (58%) compared to the average for Italy (38%) (Pearson's Chi-square test, P-value = 0.0005). While haplogroup U was found to have approximately the same frequency as other Italian populations, haplogroup T was 5% compared to 12% in Italy generally (7% in the North). Other differences were apparent, but sample sizes were relatively low to yield significant statistical differences.

Differences are more important when examining Y-chromosome haplogroup frequencies. R1b3 reached 52% in Ladin populations but only 31% in the general population, and also in the North (Pearson's Chi-square test, P-value = 0.0087); Figure 4. More remarkable are the differences when considering the remaining R1b3 lineages, that is, R1b(xR1b3), which account for 15% of the lineages in Ladins, but only for 1% in the general

population (Pearson's Chi-square test, P-value = 0.0001). Other haplogroups showed substantial haplogroup differences (e.g. J2) but the sample size was again too small.

Due to the availability of data for mtDNA in several Ladin communities, we were able to carry out an AMOVA analysis in order to investigate the level of population stratification in these communities. The data indicated that among-population variance is 1.09%, a value that is therefore significantly higher than the average for the Italian Peninsula (0.79%; **Table 5**).

Some interesting features were also found for Ladin populations when examined to the haplotype level. For instance, the HVS-I profile G16129A C16192T A16270G T16304C was found in four Ladins from Val Badía; this profile belongs to haplogroup U5b3f [18]. In a large in-house database of worldwide profiles (>130,000 HVS-I segments), this sequence was only found sporadically in other Italian regions and in Spain (Catalonia, Galicia, and Ibiza in the Balearic Islands). U5b3f is a minor clade of U5b3, the only haplogroup reported to date that has been found to represent the glacial refuge zone in Northern Italy and a source population for human re-colonization of the continent at the beginning of the Holocene. The study of Pala et al. [18] indicates that this lineage mainly expanded along the Mediterranean coast towards the Iberian Peninsula; one sub-clade also reached Sardinia 7000–9000 years ago. The branch observed in the Ladins is younger and could also have participated in the Mediterranean spread of U5b3f towards Iberia, given its presence in modern-day Spain. The data suggest that the U5b3f members observed in the Ladins probably predate the Ladin ethnogenesis and, given that this population has somehow become isolated from other neighboring populations, could reach a substantial frequency in some other Ladin communities, as is the case for the Val Badia. Another example is the U3 profile A16233G C16256T T16311C A16343G, which was only found in five Ladins from three different communities (Val Badia in South Tyrol, Val Badia in Brunico, and Val Gardena), while T16352C C16354T was found in six individuals from Val Badia in South Tyrol.

Diversity values in the Grecani Salentini samples were similar to those observed in other Italian regions. Moreover, they also show haplogroup frequency patters in the Y-chromosome and the mtDNA that matches well with other Italian samples. The haplogroups are typically European (**Figure 4**); given the southern location of the Grecani Salentini in the Italian Peninsula, it is noticeable that there is no evidence of North African lineages. Note

however, that at other level of phylogenetic resolution, there are signals on the Y-chromosome of North African enrichment in South Italy [29].

The North African historical legacy in South Italy and the Lucera population

We sampled 60 individuals from Lucera. This population sample showed diversity values that fell within the average of a typical Italian population, regarding the mtDNA (**Table 1**) and the Y-chromosome (**Table 2**). Additionally, at the level of haplogroup frequencies, Lucera matched well with other Italian populations (**Figure 4**).

There are two mtDNA haplogroups, namely U6 and M1, that can be considered to be of North African origin and could therefore be used to signal the documented historical input of this African region into Lucera. In our full set of samples, we observed five U6 haplotypes belonging to sub-haplogroups U6a, U6a2, and U6a4. Only one of these haplotypes was observed in Lucera. However, the other three U6 haplotypes were observed in the vicinity of the population of Messapi, and another at the tip of the Peninsula (Calabria). Regarding M1 haplotypes, we observed only two carriers in our samples sharing the same HVS-I haplotype; both were found in Trapani (West Sicily).

Therefore, while South Italy shows evidence of having female introgression from North Africa, this African influence seems not to be particularly centered in the Lucera. In the Y-chromosome, we did not observe any signal of North African introgression; at least, no more than for other regions of Italy (perhaps with the exception of Sicily [29]). This again contrast with the results of previous studies based on the Y-chromosome (but at higher level of phylogenetic resolution) where signals of North African influence were observed at this latitude of the Peninsula [29].

Discussion

A meta-analysis of Y-chromosome and mtDNA sequence data was undertaken in order to investigate patterns of genetic variation throughout Italy. Molecular indices indicated that most of the Italian samples show diversity values that are comparable to other European populations. However, some differences were shown to exist, especially in isolated Ladin populations. Regional differences were much more evident when examining haplogroup frequencies in both uniparental markers. The differences were again more remarkable for the two linguistic isolates, the Ladins and Grecani Salentini. AMOVA also

indicated the existence of significant population stratification along the length of the country, which appeared more remarkable for the Y-chromosome and for haplogroups than for haplotypes. These figures have however to be considered with caution given the different mutability of the markers being analyzed [64]; see also a discussion in [65].

Over the last few years, the interest in genetically isolated populations has increased, especially in biomedical studies, where there exists a growing interest in revealing genetic variants associated to disease. Genetic isolates generally originate as a result of group "foundation" by a small number of individuals presenting initially low variability. We have here analyzed a new sample of the Ladins, a well-known isolate from the Italian Alps. Some investigations were focused on the Ladin Romance speaking populations, distributed between Trentino, the Veneto regions and South Tirol area [20,62,63,66]. As also observed in the present study, Ladin communities show marked genetic differentiation with neighboring (non-Ladin) populations. Differences were also observed between the different Ladin groups; for instance, AMOVA analysis also indicated that the different Ladin communities show a level of population stratification that is higher than the average in the rest of Italy. These results are also consistent with the recent study by Coia et al. [67], derived from micro-geographical analysis of nine sample populations from Trentino (Eastern Italian Alps). Genetic differences between Ladin samples are most likely to be due to the limited historical gene flow existing between theses communities [20]. In this regard, it is also noticeable that, while the South Tirol populations show clear signatures of isolation, the Veneto groups presented a high degree of genetic variability [68].

Conversely, the Grecani Salentini also showed signatures of genetic isolation when compared to other Italian populations, but the differences are not as marked as observed for the Ladins. The differences with respect to neighboring Italian populations were not evident when observing individual haplotypes (as occurs with the Ladins), but were clearer when considering haplogroup frequencies (**Figure 4**); for instance, haplogroup U reaches 28% in the Grecani while it is about 21% in the general population. Larger sample sizes are needed in order to gather more signatures about the demographic past of this population. Thus, the Ladins show a more distinctive pattern than the Grecani Salentini, which is to be expected given that not only is the Ladin population a linguistic isolate, but also that these communities are confined to isolated geographical areas of the Alps.

Apart from the regional and local genetic differences observed in Italy, it is also worth examining global genetic patterns along the length of continental Italy.

Geographical clines of Y-chromosome haplogroups in Europe have been previously reported in the literature [11]; these patterns have found support in archaeological and linguistic evidence. In the Italian peninsula, the Y-chromosome variation also shows a clinal pattern along the North–South axis; the Mesolithic haplogroup R1*(xR1a1) shows higher frequency in the North while the Neolithic haplogroup J2-M172 is superposed to this Mesolithic strata with frequency patterns running in the opposite direction [12,69]. The results of the present study agreed with these earlier findings. Thus, for instance, R1b3 reached 31% in the North, 16% in the Center, and 14% in the South. Frequency of J2 was found to be 9% in the North, 37% in the Center, and 22% in the South (average in Italy: 14.5%). Haplogroup J2 is widely believed to be associated with the spread of agriculture from Mesopotamia. The main spread of J2 into the Mediterranean area is thought to have coincided with the expansion of agricultural populations during the Neolithic period. As reported by Di Giacomo et al. [10], haplogroup J "…constitutes not only the signature of a single wave-of-advance from the Levant but, to a greater extent, also of the expansion of the Greek world, with an accompanying novel quota of genetic variation produced during its demographic growth…"; also that "…in the central and west Mediterranean, the entry of J chromosomes may have occurred mainly by sea, i.e., in the south–east of both Spain and Italy...". J2-M12 is almost totally represented by its sublineage J2-M102, which shows frequency peaks in both the southern Balkans and north-central Italy (14%; [11]). J2-M67 is most frequent in the Caucasus, and J2-M92 indicates affinity between Anatolia and southern Italy (21.6%; [11]). For the J1-M170 clade, the peaks of J1-M267 are in the Levant and in northern Africa, and it is closely associated to the diffusion of the Arab people, dropping abruptly outside of this area (including Anatolia and the Iberian peninsula), even if it shows an appreciable percentage in Sicily [70].

Latitudinal clinal frequency patterns are also observed for the mtDNA haplogroups mirroring those of the Y-chromosome. As reported by Richards et al. [36], haplogroups H, K, T*, T2, W, and X are the major contributors to the Late Upper Paleolithic, and the central-Mediterranean region has the greatest Middle Upper Paleolithic component outside the Caucasus. In agreement with the Y-chromosome, we observed that all these Paleolithic haplogroups together add to approximately 70.3% in the North, 60.8% in the Center, and

54% in the South of Italy. The opposite pattern was observed for the main mtDNA Neolithic component, represented by haplogroups J and T1, which accounted for 5.8% in the North, 10.3% in the Center, and 14.1% in the South (Italian average: 10.5%).

As early as 1934, [71], Vere Gordon Childe suggested that the indigenous communities of hunters and gatherers of the Mesolithic European cultures were replaced by communities of farmers migrating to the North from the Middle East, a process that lasted for several generations. The first stream of emigration followed the route along the continental Balkan Peninsula and the Danube, while another, slightly later, emigration spread through shipping along the coasts of the Mediterranean Sea from East to West. The latter path would fit well with the distribution of other Neolithic cultural features, such as the so-called Cardium Pottery (or Cardial Ware) [72], the ceramic decorative style that better defines the Neolithic culture. This culture entered from Greece towards the South-Center of Italy through the Adriatic Sea, carried by the same farmers that introduced, for instance, Y-chromosome haplogroup J2 at about the same frequency in Central and South Italy, but with lower introgression into the North; from here followed further Mediterranean expansions towards Iberia.

The sub-clade E3b1 (probably originating in eastern Africa) has a wide distribution in Africa, Near East and Europe. This haplogroup reaches a frequency of 8% in the North and Center and slightly higher in the South, 11% (**Figure 1**). It has also been argued that the European distribution of E3b1 is compatible with the Neolithic demic diffusion of agriculture [13]; thus, two sub-clades, E3b1a- M78 and E3b1c-M123 present a higher occurrence in Anatolia, the Balkans and the Italian peninsula. Another sub-clade, E3b1b-M81 is associated with the Berber populations and is commonly found in regions that have had historical gene flow with Northern Africa, such as the Iberian peninsula [73,74]–[75-77], including the Canary Islands [74], and Sicily [70,78]; the absence of microsatellite variation suggests a very recent arrival from North Africa [79]. If we assume that all E3b1 represents the only Y-chromosome African component in Italy and L and U6 lineages the African mtDNA, the African component in Italy is higher for the Y-chromosome (8–11%) than for mtDNA (1–2%). The origin of sub-Saharan African mtDNAs in Europe (including Italian samples) has been recently investigated by Cerezo et al. [80]; the results indicate that a significant proportion of these lineages could have arrived in Italy more than 10,000 years ago;

therefore, their presence in Europe does not necessarily date to the time of the Roman Empire, the Atlantic slave trade or to modern migration.

The Northern African influence in the Italian Peninsula is evidenced by the presence of Northern African Y chromosome haplogroups (E1-M78) in three geographically close samples across the southern Apennine mountains: East Campania, Northwest Apulia and Lucera [29]. The Lucera sample analyzed in the present study did not however show a higher impact from North Africa than for other areas from southern Italy [29].

Finally, in agreement with uniparental markers, analysis of AIMs as carried out in the present study indicated that Italy has a main European ancestry, and shows a very minor sub-Saharan African component that is, however, slightly higher than non-Mediterranean Europe.

The present study represents the largest meta-analysis carried out to date for the Italian peninsula. We observed that the Y-chromosome and the mtDNA retain the imprint of the major ancestral events occurring in Italy; however, the Y-chromosome shows more marker regional differences than does the mtDNA. It is difficult to infer what proportion of these differences can be attributed not only exclusively to gender demographic differences, but also to the fact that both markers were analyzed to different levels of molecular resolution. Italy shows clines of variation attributable to the demographic movements of the first Paleolithic settlements, posteriorly modeled by the Mesolithic and, to a lesser extent, Neolithic farmers. Regional differences arose with time, which are more notable in linguistic isolates, such as the Ladin populations, and to a minor extent, the Grecani Salentini.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement number 290344, and the Ministerio de Ciencia e Innovación (SAF2008-02971 and SAF2011-26983) (AS). CC and FB were partially funded by the British Academy for the project "The Greeks in the West: the genetic legacy of the colonisation in South Italy and Sicily".

References

1. Cunliffe B (2001) The Oxford Illustrated History of Prehistoric Europe: Oxford University Press.
2. Buti GG (1974) Preistoria e storia delle regioni d'Italia: Sansoni Università.
3. Devoto G (1977) Gli antichi italici: Firenze, Vallecchi.
4. Pallottino M (1981) Genti e culture dell'Italia preromana. Jouvence: 136.
5. Stenico M, Nigro L, Barbujani G (1998) Mitochondrial lineages in Ladin-speaking communities of the eastern Alps. Proc R Soc Lond B 265: 555-561.
6. Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, et al. (2007) Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. Am J Hum Genet 80: 759-768.
7. Brisighelli F, Capelli C, Álvarez-Iglesias V, Onofri V, Paoli G, et al. (2009) The Etruscan timeline: A recent Anatolian connection. Eur J Hum Genet 17: 693-696.
8. Francalacci P, Bertranpetit J, Calafell F, Underhill PA (1996) Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. Am J Phys Anthropol 100: 443-460.
9. Onofri V, Alessandrini F, Turchi C, Fraternale B, Buscemi L, et al. (2007) Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis. Int J Legal Med 121: 234-237.
10. Di Giacomo F, Luca F, Anagnou N, Ciavarella G, Corbo RM, et al. (2003) Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. Mol Phylogenet Evol 28: 387-395.
11. Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, et al. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: A Y chromosome perspective. Science 290: 1155-1159.
12. Capelli C, Brisighelli F, Scarnicci F, Arredi B, Caglia A, et al. (2007) Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. Mol Phylogenet Evol 44: 228-239.
13. Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, et al. (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: Inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet 74: 1023-1034.
14. Trombetta B, Cruciani F, Sellitto D, Scozzari R (2011) A new topology of the human Y chromosome haplogroup E1b1 (E-P2) revealed through the use of newly characterized binary polymorphisms. PLoS ONE 6: e16073.
15. Francalacci P, Sanna D (2008) History and geography of human Y-chromosome in Europe: A SNP perspective. J Anthropol Sci 86: 59-89.
16. Di Gaetano C, Cerutti N, Crobu F, Robino C, Inturri S, et al. (2009) Differential Greek and northern African migrations to Sicily are supported by genetic evidence from the Y chromosome. Eur J Hum Genet 17: 91-99.
17. Brigaglia M (1989) Storia della Sardegna. Edizioni Della Torre.
18. Pala M, Achilli A, Olivieri A, Kashani BH, Perego UA, et al. (2009) Mitochondrial haplogroup U5b3: A distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. Am J Hum Genet 84: 814-821.

19. Montanini L, Regna-Gladin C, Eoli M, Albarosa R, Carrara F, et al. (2005) Instability of mitochondrial DNA and MRI and clinical correlations in malignant gliomas. J Neurooncol 74: 87-89.
20. Thomas MG, Barnes I, Weale ME, Jones AL, Forster P, et al. (2008) New genetic evidence supports isolation and drift in the Ladin communities of the South Tyrolean Alps but not an ancient origin in the Middle East. Eur J Hum Genet 16: 124-134.
21. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, et al. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75: 910-918.
22. Torroni A, Bandelt H-J, Macaulay V, Richards M, Cruciani F, et al. (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69: 844-852.
23. Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintáns B, Zarrabeitia MT, et al. (2009) New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS One 4: e5112.
24. Gómez-Carballa A, Olivieri A, Behar DM, Achilli A, Torroni A, et al. (2012) Genetic continuity in the Franco-Cantabrian region: New clues from autochthonous mitogenomes. PLoS ONE 7: e32851.
25. De Concini W (2003) Gli altri d'Italia : minoranze linguistiche allo specchio Pergine Valsugana (TN).
26. Castagnetti A (2004) Storia del Trentino. (L'età Medievale). 3 Vol, Bologna, Il Mulino, 2000-2005.
27. De Concini W (1997) Gli altri delle Alpi. Minoranze linguistiche dell'arco alpino italiano. Trento, Grafiche Artigianelli.
28. Carducci L (1993) Storia del Salento. La terra de'Otranto dale origini ai primi del cinquecento. Galatina: Congedo.
29. Capelli C, Onofri V, Brisighelli F, Boschi I, Scarnicci F, et al. (2009) Moors and Saracens in Europe: Estimating the medieval North African male legacy in southern Europe. Eur J Hum Genet 17: 848-852.
30. Norman D (1975) The Arabs and Medieval Europe. London, UK Longmann Group Limited.
31. Brisighelli F, Blanco-Verea A, Boschi I, Garagnani P, Pascali VL, et al. (2011) Global patterns of Y-STR variation in Italy. Forensic Sci Int Genet in press; 10.1016/j.fsigen.2012.03.003.
32. Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16: 1215.
33. Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int Genet 1: 44-55.
34. Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, et al. (1981) Sequence and organization of the human mitochondrial genome. Nature 290: 457-465.
35. Quintáns B, Álvarez-Iglesias V, Salas A, Phillips C, Lareu MV, et al. (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140: 251-257.
36. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251-1276.
37. Brión M, Sobrino B, Blanco-Verea A, Lareu MV, Carracedo Á (2005) Hierarchical analysis of 30 Y-chromosome SNPs in European populations. Int J Legal Med 119: 10-15.

38. YCC (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12: 339-348.
39. Sánchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, et al. (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 27: 13-24.
40. Amigo J, Salas A, Phillips C (2011) ENGINES: exploring single nucleotide variation in entire human genomes. BMC Bioinformatics 12: 105.
41. Amigo J, Salas A, Phillips C, Carracedo Á (2008) SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. BMC Bioinformatics 9: 428.
42. Nei N (1987) Molecular evolutionary genetics: New York: Columbia University Press.
43. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437-460.
44. Tajima F (1993) Unbiased estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 10: 677-688.
45. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.
46. Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Res 10: 564-567.
47. Loogväli E-L, Roostalu U, Malyarchuk BA, Derenko MV, Kivisild T, et al. (2004) Disuniting uniformity: A pied cladistic canvas of mtDNA haplogroup H in Eurasia. Mol Biol Evol 21: 2012-2021.
48. Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, et al. (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64: 232-249.
49. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30: E386-394.
50. Salas A, Yao Y-G, Macaulay V, Vega A, Carracedo A, et al. (2005) A critical reassessment of the role of mitochondria in tumorigenesis. PLoS Med 2: e296.
51. Salas A, Prieto L, Montesino M, Albarrán C, Arroyo E, et al. (2005) Mitochondrial DNA error prophylaxis: Assessing the causes of errors in the GEP'02-03 proficiency testing trial. Forensic Sci Int 148: 191-198.
52. Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt H-J (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335: 891-899.
53. Salas A, Bandelt H-J, Macaulay V, Richards MB (2007) Phylogeographic investigations: The role of trees in forensic genetics. Forensic Sci Int 168: 1-13.
54. Yao Y-G, Salas A, Logan I, Bandelt H-J (2009) mtDNA data mining in GenBank needs surveying. Am J Hum Genet 85: 929-933; author reply 933.
55. Bandelt H-J, Salas A, Lutz-Bonengel S (2004) Artificial recombination in forensic mtDNA population databases. Int J Legal Med 118: 267-273.
56. Bandelt H-J, Salas A, Bravi CM (2004) Problems in FBI mtDNA database. Science 305: 1402-1404.
57. Bandelt H-J, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71: 1150-1160.
58. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, et al. (2006) The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science 314: 1767-1770.
59. Salas A, Richards M, De la Fé T, Lareu MV, Sobrino B, et al. (2002) The making of the African mtDNA landscape. Am J Hum Genet 71: 1082-1111.

60. Catelli ML, Álvarez-Iglesias V, Gomez-Carballa A, Mosquera-Miguel A, Romanini C, et al. (2011) The impact of modern migrations on present-day multi-ethnic Argentina as recorded on the mitochondrial DNA genome. BMC Genet 12: 77.
61. Gómez-Carballa A, Ignacio-Veiga A, Álvarez-Iglesias V, Pastoriza-Mourelle A, Ruiz Y, et al. (2012) A melting pot of multicontinental mtDNA lineages in admixed Venezuelans. Am J Phys Anthropol 147: 78-87.
62. Vernesi C, Fuselli S, Castri L, Bertorelle G, Barbujani G (2002) Mitochondrial diversity in linguistic isolates of the Alps: A reappraisal. Hum Biol 74: 725-730.
63. Pichler I, Mueller JC, Stefanov SA, De Grandi A, Beu Volpato C, et al. (2006) Genetic structure in contemporary South Tyrolean isolated populations revealed by analysis of Y-Chromosome, mtDNA, and Alu polymorphisms. Hum Biol 78: 441-464.
64. Meirmans PG, Hedrick PW (2011) Assessing population structure: F(ST) and related measures. Mol Ecol Resour 11: 5-18.
65. Brisighelli F, Blanco-Verea A, Boschi I, Garagnani P, Pascali VL, et al. (2012) Patterns of Y-STR variation in Italy. Forensic Sci Int Genet in press.
66. Stenico M, Nigro L, Bertorelle G, Calafell F, Capitanio M, et al. (1996) High Mitochondrial Sequence Diversity in Linguistic Isolates of the Alps. Am J Hum Genet 59: 1363-1375.
67. Coia V, Boschi I, Trombetta F, Cavulli F, Montinaro F, et al. (2012) Evidence of high genetic variation among linguistically diverse populations on a micro-geographic scale: a case study of the Italian Alps. J Hum Genet 57: 254-260.
68. Destro Bisol G, Anagnostou P, Batini C, Battaggia C, Bertoncini S, et al. (2008) Italian isolates today: geographic and linguistic factors shaping human biodiversity. J Anthropol Sci 86: 179-188.
69. Busby GB, Brisighelli F, Sánchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, et al. (2012) The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci 279: 884–892.
70. Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, et al. (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: Inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet 74: 1023-1034.
71. Childe VG (1934) New light on the most ancient East: The oriental Prelude to European prehistory: London: K Paul, Trench, Trubner & Co.
72. King R, Underhill P (2002) Congruent distribu- tion of Neolithic painted pottery and ceramic figurines with Y-chromosome lineages. Antiquity 76: 707–714.
73. Maca-Meyer N, Sanchez-Velasco P, Flores C, Larruga JM, Gonzalez AM, et al. (2003) Y chromosome and mitochondrial DNA characterization of Pasiegos, a human isolate from Cantabria (Spain). Ann Hum Genet 67: 329-339.
74. Flores C, Maca-Meyer N, Perez JA, Gonzalez AM, Larruga JM, et al. (2003) A predominant European ancestry of paternal lineages from Canary Islanders. Ann Hum Genet 67: 138-152.
75. Goncalves R, Freitas A, Branco M, Rosa A, Fernandes AT, et al. (2005) Y-chromosome lineages from Portugal, Madeira and Acores record elements of Sephardim and Berber ancestry. Ann Hum Genet 69: 443-454.
76. Alonso S, Flores C, Cabrera V, Alonso A, Martin P, et al. (2005) The place of the Basques in the European Y-chromosome diversity landscape. Eur J Hum Genet 13: 1293-1302.
77. Beleza S, Gusmão L, Lopes A, Alves C, Gomes I, et al. (2006) Micro-phylogeographic and demographic history of Portuguese male lineages. Ann Hum Genet 70: 181-194.

78. Cruciani F, La Fratta R, Santolamazza P, Sellitto D, Pascone R, et al. (2004) Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. Am J Hum Genet 74: 1014-1022.
79. Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, et al. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68: 1019-1029.
80. Cerezo M, Achilli A, Olivieri A, Perego UA, Gómez-Carballa A, et al. (2012) Reconstructing ancient mitochondrial DNA links between Africa and Europe. Genome Res in press.
81. Amigo J, Phillips C, Salas A, Carracedo A (2009) Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes. BMC Bioinformatics 10: S5.
82. Turchi C, Buscemi L, Giacchino E, Onofri V, Fendt L, et al. (2009) Polymorphisms of mtDNA control region in Tunisian and Moroccan populations: An enrichment of forensic mtDNA databases with Northern Africa data. Forensic Sci Int Genet 3: 166-172.
83. Messina F, Scorrano G, Labarga CM, Rolfo MF, Rickards O Mitochondrial DNA variation in an isolated area of Central Italy. Ann Hum Biol 37: 385-402.
84. Babalini C, Martínez-Labarga C, Tolk HV, Kivisild T, Giampaolo R, et al. (2005) The population history of the Croatian linguistic minority of Molise (southern Italy): A maternal view. Eur J Hum Genet 13: 902-912.
85. Falchi A, Giovannoni L, Calo CM, Piras IS, Moral P, et al. (2006) Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms. J Hum Genet 51: 9-14.
86. Verginelli F, Donati F, Coia V, Boschi I, Palmirotta R, et al. (2003) Variation of the hypervariable region-1 of mitochondrial DNA in central-eastern Italy. J Forensic Sci 48: 443-444.
87. Bini C, Ceccardi S, Luiselli D, Ferri G, Pelotti S, et al. (2003) Different informativeness of the three hypervariable mitochondrial DNA regions in the population of Bologna (Italy). Forensic Sci Int 135: 48-52.
88. Tagliabracci C, Turchi C, Buscemi L, Sassaroli C (2001) Polymorphism of the mitochondrial DNA control region in Italians. Int J Legal Med 114: 224-228.
89. Ottoni C, Martinez-Labarga C, Vitelli L, Scano G, Fabrini E, et al. (2009) Human mitochondrial DNA variation in Southern Italy. Ann Hum Biol 36: 785-811.
90. Rose G, Longo T, Maletta R, Passarino G, Bruni AC, et al. (2008) No evidence of association between frontotemporal dementia and major European mtDNA haplogroups. Eur J Neurol 15: 1006-1008.
91. Ottoni C, Martinez-Labarga C, Loogvali EL, Pennarun E, Achilli A, et al. (2009) First genetic insight into Libyan Tuaregs: a maternal perspective. Ann Hum Genet 73: 438-448.
92. Cali F, Le Roux MG, D'Anna R, Flugy A, De Leo G, et al. (2001) MtDNA control region and RFLP data for Sicily and France. Int J Legal Med 114: 229-231.
93. Forster P, Cali F, Rohl A, Metspalu E, D'Anna R, et al. (2002) Continental and subcontinental distributions of mtDNA control region types. Int J Legal Med 116: 99-108.
94. Vona G, Ghiani ME, Calo CM, Vacca L, Memmi M, et al. (2001) Mitochondrial DNA sequence analysis in Sicily. Am J Hum Biol 13: 576-589.
95. Di Rienzo A, Wilson AC (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci U S A 88: 1597-1601.

96. Ferri G, Ceccardi S, Lugaresi F, Bini C, Ingravallo F, et al. (2008) Male haplotypes and haplogroups differences between urban (Rimini) and rural area (Valmarecchia) in Romagna region (North Italy). Forensic Sci Int 175: 250-255.
97. Ferri G, Alu M, Corradini B, Radheshi E, Beduschi G (2009) Slow and fast evolving markers typing in Modena males (North Italy). Forensic Sci Int Genet 3: e31-33.
98. Contu D, Morelli L, Santoni F, Foster JW, Francalacci P, et al. (2008) Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: Inference for association scans. PLoS One 3: e1430.
99. Pichler I, Mueller JC, Stefanov SA, De Grandi A, Volpato CB, et al. (2006) Genetic structure in contemporary south Tyrolean isolated populations revealed by analysis of Y-chromosome, mtDNA, and Alu polymorphisms. Hum Biol 78: 441-464

Table 1. Diversity indices computed for different Italian regions based on HVS-I data (sequence segment 16090-16365)

Population	*Region*	*Pop ID*	*Reference*	*N*	*k*	*k/n*	*S*	*h*	Π	*M*
Liguri	NW	1	p.s.	50	40	0.8	53	0.962±0.021	0.01426±0.0145	4.875
Torino	NW	5	(Turchi et al. 2009)	50	45	0.9	49	0.993±0.007	0.01483±0.0011	5.056
Ladin	NE	2, 13	p.s. (Pichler et al. 2006; Stenico et al. 1996; Thomas et al. 2008; Vernesi et al. 2002)	504	170	0.3	106	0.960±0.005	0.01251±0.0004	4.252
Pavia	NE	6	(Turchi et al. 2009)	47	35	0.7	44	0.969±0.017	0.01316±0.0012	4.502
Udine	NE	3	p.s.	51	43	0.8	54	0.989±0.008	0.01437±0.0009	4.858
Arezzo/Chiusi	CW	1	(Brisighelli et al. 2009)	14	14	1	22	1.000±0.027	0.01488±0.0129	5.088
Casentino	CW	15	(Achilli et al. 2007)	122	77	0.6	167	0.979±0.007	0.02409±0.0082	8.190
Collevecchio/Magliano Sabino	CW	3	(Brisighelli et al. 2009)	12	11	0.9	14	0.985±0.040	0.01201±0.0015	4.106
Elba	CW	2	(Brisighelli et al. 2009)	16	6	0.4	11	0.683±0.120	0.00853±0.0017	2.908
Firenze	CW	9	(Turchi et al. 2009)	48	40	0.8	54	0.980±0.014	0.01332±0.0012	4.556
Jenne	CW	22	(Messina et al.)	103	34	0.3	47	0.834±0.036	0.01006±0.0360	3.440
Latini	CW	5	p.s.	48	29	0.6	35	0.902±0.039	0.01003±0.0010	3.429
Latium	CW	20	(Babalini et al. 2005)	52	37	0.7	48	0.959±0.019	0.01313±0.0014	4.492
Murlo	CW	16	(Achilli et al. 2007)	86	60	0.7	68	0.976±0.010	0.01327±0.0009	4.524
Roma	CW	12	(Turchi et al. 2009)	58	49	0.8	55	0.987±0.008	0.01433±0.0011	4.901
Terni	CW	11	(Turchi et al. 2009)	29	20	0.7	33	0.941±0.034	0.01201±0.0014	4.108
Tuscany	CW	4	(Brisighelli et al. 2009; Falchi et al. 2006; Francalacci et al. 1996)	127	86	0.7	77	0.982±0.007	0.01305±0.0075	4.464
Vallepietra	CW	21	(Messina et al.)	21	8	0.4	17	0.871±0.044	0.01281±0.0014	4.381
Volterra	CW	14	(Achilli et al. 2007)	114	57	0.5	62	0.955±0.013	0.01193±0.0007	4.057
Abruzzo	CE	17	(Babalini et al. 2005; Verginelli et al. 2003)	61	53	0.8	62	0.990±0.007	0.01500±0.0010	5.131
Ancona	CE	10	(Turchi et al. 2009)	73	55	0.7	59	0.963±0.017	0.01379±0.0010	4.717
Bologna	CE	7	(Bini et al. 2003; Turchi et al. 2009)	146	79	0.5	64	0.970±0.008	0.01250±0.0006	4.278
Centre East	CE	23	(Tagliabracci et al. 2001)	83	62	0.7	60	0.974±0.012	0.01352±0.0009	4.625
Croatian Italians	CE	19	(Babalini et al. 2005)	41	28	0.7	46	0.970±0.015	0.01524±0.0017	5.213
Modena	CE	8	(Turchi et al. 2009)	44	33	0.7	43	0.958±0.023	0.01139±0.0012	3.895
Molise	CE	18	(Babalini et al. 2005)	62	41	0.6	58	0.938±0.025	0.01260±0.0013	4.309
Piceni	CE	4	p.s.	53	43	0.8	56	0.985±0.009	0.01306±0.0011	4.414
Belvedere	SW	10	p.s.	50	41	0.8	44	0.980±0.013	0.01320±0.0010	4.532
Calabria	SW	27	(Ottoni et al. 2009b; Rose et al. 2008)	389	213	0.5	128	0.983±0.003	0.01521±0.0004	5.203
Campania	SW	30	(Babalini et al. 2005)	48	41	0.8	59	0.980±0.014	0.01519±0.0014	5.166
Catania	SW	11	p.s.	40	35	0.9	45	0.990±0.010	0.01460±0.0012	4.979
Sicily	SW	28	(Cali et al. 2001; Forster et al. 2002; Ottoni et al. 2009a; Richards et al. 2000; Vona et al. 2001)	558	240	0.4	125	0.958±0.006	0.01289±0.0004	4.343
Trapani	SW	12	p.s.	40	30	0.7	36	0.977±0.013	0.01313±0.0013	4.465
Apulia	SE	26	(Babalini et al. 2005)	26	24	0.9	43	0.991±0.015	0.01550±0.0022	5.304
Basilicata	SE	25	(Ottoni et al. 2009b)	92	65	0.7	70	0.983±0.007	0.01290±0.0008	4.428
Grecani Salentini	SE	8	p.s.	47	37	0.8	44	0.989±0.007	0.01310±0.0011	4.480
Lucera	SE	6	p.s.	60	42	0.7	55	0.976±0.011	0.01345±0.0011	4.586
Messapi	SE	9	p.s.	53	38	0.7	49	0.973±0.014	0.01579±0.0010	5.401
Sanniti	SE	7	p.s.	50	41	0.8	49	0.988±0.008	0.01420±0.0013	4.843
Sardinia	–	29	(Di Rienzo and Wilson 1991; Falchi et al. 2006; Richards et al. 2000)	351	171	0.4	98	0.950±0.009	0.01183±0.0004	4.033
Geographical region										
North Italy	–	–		702	267	0.4	126	0.963±0.004	0.01282±0.0004	4.295
Centre Italy	–	–		1413	500	0.4	216	0.958±0.004	0.01243±0.0002	4.113
South Italy	–	–		1453	569	0.4	183	0.973±0.002	0.01368±0.0002	4.541
West Italy (without Sicily)	–	–		1437	578	0.4	232	0.969±0.003	0.01315±0.0002	4.405
West Italy (with Sicily)	–	–		2075	709	0.3	236	0.963±0.003	0.01260±0.0002	4.133
East Italy	–	–		1493	520	0.3	165	0.964±0.003	0.01277±0.0002	4.200

NW = north-west; NE = north-east; CW = centre-west; CE = centre-east; SW = south-west; SE = south-east; *N* = sample size; k = number of different haplotypes; S= segregating sites; h = haplotype diversity; π = nucleotide diversity; M = average number of nucleotide differences

Table 2. Diversity indices computed for different Italian regions based on Y-SNPs. Codes are as in Table 1.

Population	*Region*	*Reference*	*N*	*k*	*k/n*	*Gene Diversity*
Liguria	NW	Present study	46	9	0.19	0.7662±0.0502
Ladin	NE	(Capelli et al. 2007)	34	6	0.17	0.5348±0.0979
Udine	NE	Present study	47	10	0.21	0.7761±0.0441
Central Tuscany	CW	(Capelli et al. 2007)	40	8	0.20	0.7397±0.0616
Elba Island	CW	(Capelli et al. 2007)	94	7	0.07	0.6742±0.0445
Latini	CW	Present study	44	11	0.25	0.8256±0.0395
Latium	CW	(Capelli et al. 2007)	43	9	0.20	0.8026±0.0388
Tuscany-Latium border	CW	(Capelli et al. 2007)	76	7	0.09	0.7554±0.0350
Central Marche	CE	(Capelli et al. 2007)	59	7	0.11	0.7294±0.0364
Marche	CE	(Onofri et al. 2007)	162	13	0.08	0.8489±0.0152
Marche-Appennine	CE	(Capelli et al. 2007)	25	7	0.28	0.8033±0.0514
Modena	CE	(Ferri et al. 2008)	62	8	0.12	0.5320±0.0743
Piceni	CE	Present study	38	9	0.23	0.8208±0.0450
Rimini-Val Marecchia	CE	(Ferri et al. 2009)	163	12	0.35	0.6990±0.0308
Belvedere	SW	Present study	27	9	0.33	0.8547±0.0477
East Campania	SW	(Capelli et al. 2007)	46	7	0.15	0.6870±0.0618
Sicily	SW	Present study	57	12	0.21	0.8327±0.0311
West Campania	SW	(Capelli et al. 2007)	80	10	0.12	0.8446±0.0224
West Calabria	SW	(Capelli et al. 2007)	57	7	0.12	0.7525±0.0307
Sanniti	SE	Present study	30	10	0.33	0.8644±0.0409
Grecani Salentini	SE	Present study	47	7	0.14	0.8122±0.0242
Lucera	SE	(Capelli et al. 2009)	60	9	0.15	0.8365±0.0236
Messapi	SE	(Capelli et al. 2007)	49	9	0.18	0.8529±0.0237
Sardinia		(Contu et al. 2008)	336	14	0.04	0.8098±0.0136
Geographical region						
North Italy	–	–	127	14	0.11	0.8400±0.0189
Centre Italy	–	–	806	21	0.03	0.8870±0.0053
South Italy	–	–	453	20	0.04	0.8909±0.0060
West Italy (without Sicily)	–	–	553	17	0.03	0.8567±0.0094
West Italy (with Sicily)	–	–	610	20	0.03	0.8705±0.0078
East Italy	–	–	776	22	0.02	0.9034±0.0037

Table 3. Number of haplotypes shared between different continental regions and Italian samples.

	Northeast Africa	**Northwest Africa**	**Europe**	**Italy**	**Near East**	**Sub-Sahara**
North East Africa	318	83	114	88	109	79
North West Africa	–	772	293	179	186	157
Europe	–		3854	543	496	125
Italy	–	–	–	1289	306	74
Near East	–	–	–	–	1674	154
Sub-Sahara	–	–	–	–	–	2635

Table 4. Admixture proportions of Italian population and different continental regions based on mtDNA HVS-I sequences.

Region	P_0	95% CI (P_0)	P_1	95% CI (P_1)	P_2	95% CI (P_2)
5 regions						
Africa NE	0.1845	0.1084-0.2605	0.1867	0.1103-0.2630	0.1947	0.1171-0.2723
Africa NW	0.2195	0.1384-0.3007	0.2230	0.1414-0.3046	0.2214	0.1401-0.3028
Europe	0.3550	0.2612-0.4488	0.3332	0.2408-0.4256	0.3286	0.2366-0.4207
Near East	0.2292	0.1469-0.3116	0.2383	0.1548-0.3218	0.2323	0.1495-0.3151
Sub-Sahara	0.0117	-0.0094-0.0328	0.0188	-0.0078-0.0454	0.0230	-0.0064-0.0523
4 regions						
Europe	0.4400	0.4205-0.4595	0.4124	0.3931-0.4317	0.4063	0.3871-0.4256
Near East	0.2770	0.2595-0.2946	0.2918	0.2740-0.3096	0.2900	0.2722-0.3077
North Africa	0.2667	0.2493-0.2840	0.2708	0.2534-0.2882	0.2727	0.2553-0.2902
Sub-Sahara	0.0163	0.0114-0.0213	0.0250	0.0188-0.0311	0.0310	0.0242-0.0378

Table 5. AMOVA analysis of main Italian regions (Permutations: 20000; *P*-value<0.0000) for the mtDNA control region data and the Y-chromosome STRs and SNPs. Sardinians were not included in the analysis. References for population samples are given in Table S2.

	All populations (%)	North vs Centre vs South (%)	West vs East (%)
HAPLOTYPES			
mtDNA (48 populations)			
Among pops	0.79	0	0
Within pops	99.21	99.25	99.21
Among pops within groups	–	0.75	0.79
Y-chromosome (15 populations)			
Among pops	2.34	1.18	0
Within pops	97.66	97.32	97.85
Among pops within groups	–	1.50	2.15
HAPLOGROUPS			
mtDNA (19 populations)			
Among pops	1.17	0.36	0
Within pops	98.83	98.72	98.83
Among pops within groups	–	0.92	1.17
Y-chromosome (24 populations)			
Among pops	13.92	0.07	0.83
Within pops	86.08	86.06	85.74
Among pops within groups	–	13.87	13.44

Table 6. Mitochondrial DNA molecular diversity values of different Ladin populations

Ladin	***REF***	***N***	***k***	***k/n***	***S***	***h***	**Π**	***M***
Val Badia	p.s (Thomas et al. 2008)	97	55	0.6	60	0.958±0.012	0.01206±0.0008	4.101
Upper Val Venosta	(Pichler et al. 2006; Thomas et al. 2008)	108	47	0.4	53	0.944±0.014	0.01067±0.0008	3.648
Lower Val Venosta	(Pichler et al. 2006; Thomas et al. 2008)	107	49	0.4	51	0.955±0.012	0.01220±0.0007	4.171
Val Gardena	(Stenico et al. 1996; Thomas et al. 2008)	56	27	0.5	42	0.906±0.027	0.01216±0.0011	4.158
Val Pusteria	(Pichler et al. 2006)	37	14	0.4	22	0.899±0.029	0.00981±0.0010	3.354
Val Isarco	(Pichler et al. 2006)	34	19	0.5	29	0.961±0.015	0.01111±0.0007	3.799
Colle S. Lucia	(Stenico et al. 1996; Vernesi et al. 2002)	30	17	0.8	33	0.947±0.022	0.01885±0.0010	6.448

Figures

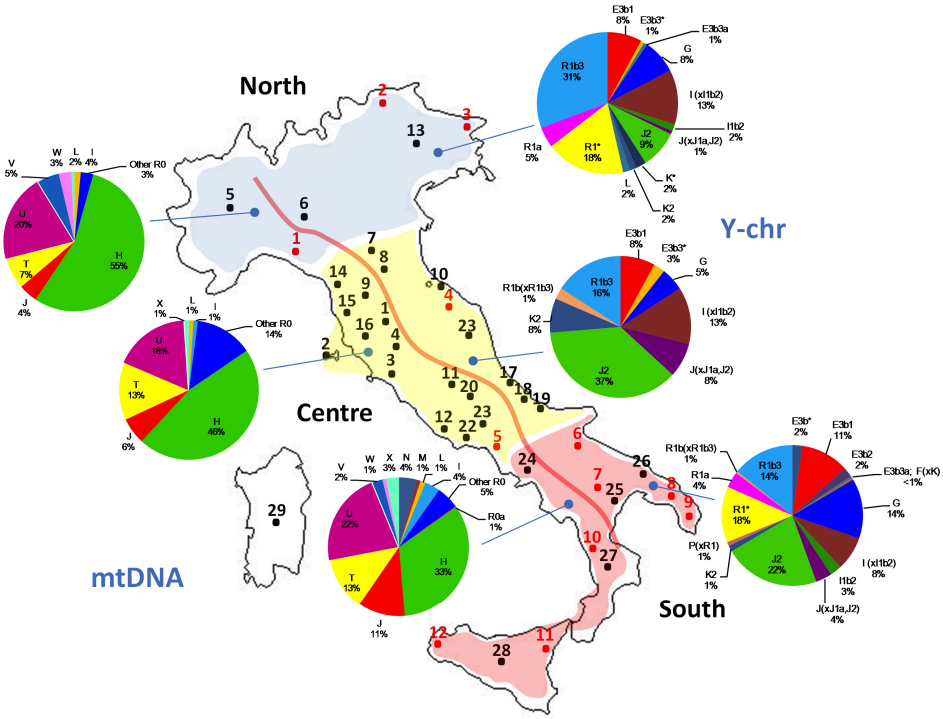


Figure1. Map showing the location of the samples analyzed in the present study and those collected from the literature (see Table 1). Pie charts on the left display the distribution of mtDNA haplogroup frequencies, and the ones on the right the Y-chromosome haplogroup frequencies.

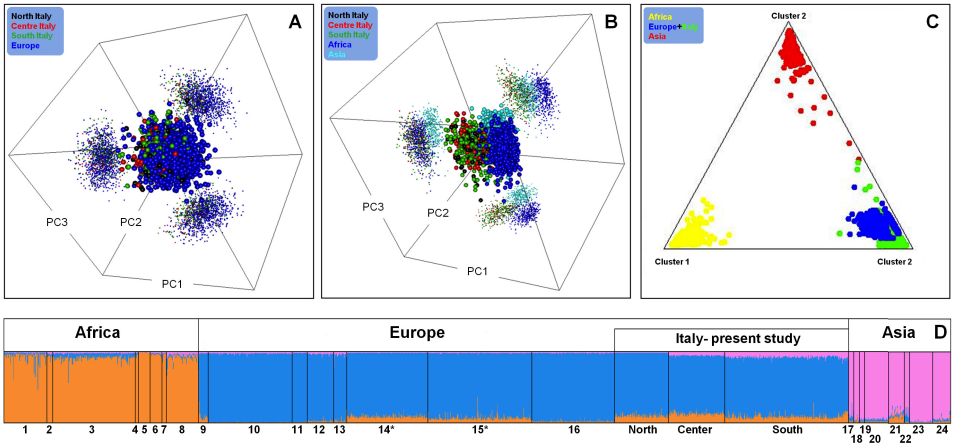
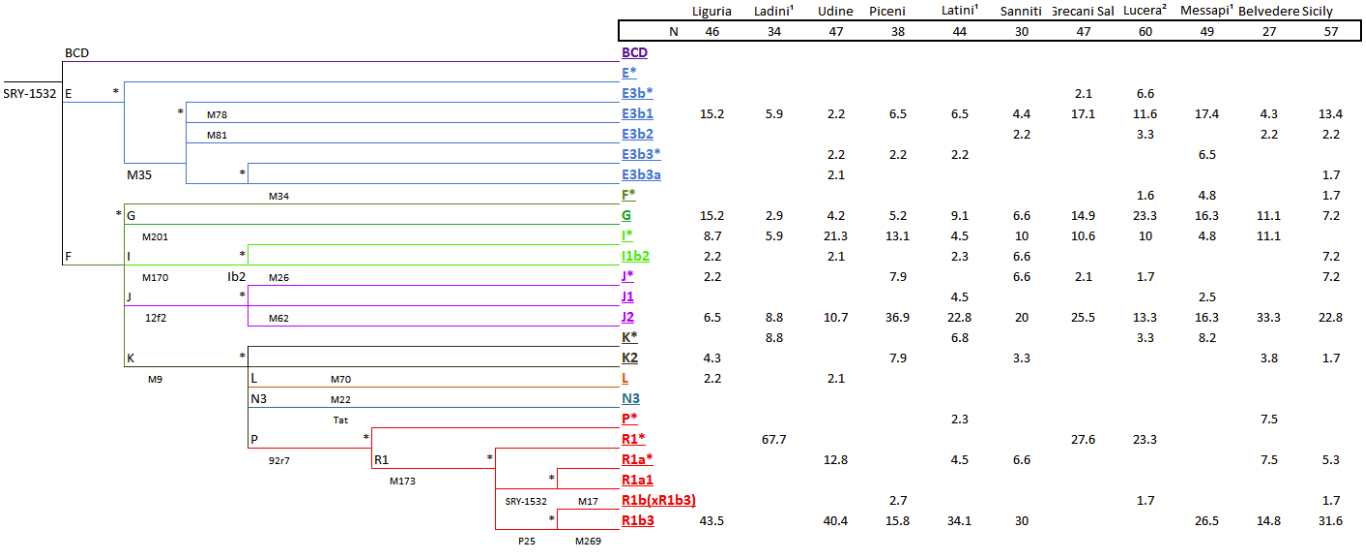


Figure 2. Analysis of AIMs in Italian populations *versus* other continental population groups. (A) PCA of Italian populations divided in main regions North, Centre and South (as analyzed in the present study) and other European populations; (B) The same Italian populations plus sub-Saharan African, and Asian ones; (C) Triangle plot as obtained using STRUCTURE of Italian, European, sub-Saharan, and Asian populations; (D) Bar plot of ancestral membership values as obtained using STRUCTURE of the same populations used in (C). Population codes: 1: Angola; 2: Kenya-Bantu NE; 3: Mozambique; 4: Namibia-San; 5: Nigeria-Yoruba; 6: Senegal-Mandenka; 7: South Africa-Bantu; 8: Uganda; 9: Britain; 10: Denmark; 11: French; 12: Germany; 13: Ireland; 14*: NW Spain; 15*: Portugal; 16: Slovenia; 17: China-Dai; 18: China-Daru; 19: China-Han ; 20: China-Hezhen; 21: Japanese; 22: Mongolia; 23: Taiwan; 24: Thailand. Genotypes were downloaded using (Amigo et al. 2009; Amigo et al. 2008) and belong to the CEPH panel. An asterisk indicates Mediterranean populations.



	Liguria	Ladini[1]	Udine	Piceni	Latini[1]	Sanniti	Grecani Sal	Lucera[2]	Messapi[1]	Belvedere	Sicily
N	46	34	47	38	44	30	47	60	49	27	57
BCD											
E*											
E3b*							2.1	6.6			
E3b1	15.2	5.9	2.2	6.5	6.5	4.4	17.1	11.6	17.4	4.3	13.4
E3b2						2.2		3.3		2.2	2.2
E3b3*			2.2	2.2	2.2				6.5		
E3b3a			2.1								1.7
F*								1.6	4.8		1.7
G	15.2	2.9	4.2	5.2	9.1	6.6	14.9	23.3	16.3	11.1	7.2
I*	8.7	5.9	21.3	13.1	4.5	10	10.6	10	4.8	11.1	
I1b2	2.2		2.1		2.3	6.6					7.2
J*	2.2			7.9		6.6	2.1	1.7			7.2
J1					4.5				2.5		
J2	6.5	8.8	10.7	36.9	22.8	20	25.5	13.3	16.3	33.3	22.8
K*		8.8			6.8			3.3	8.2		
K2	4.3			7.9		3.3				3.8	1.7
L	2.2		2.1								
N3											
P*					2.3					7.5	
R1*		67.7					27.6	23.3			
R1a*			12.8		4.5	6.6				7.5	5.3
R1a1											
R1b(xR1b3)				2.7				1.7			1.7
R1b3	43.5		40.4	15.8	34.1	30			26.5	14.8	31.6

[1] Capelli et al, 2007a
[2] Capelli et al, 2009

Figure 3. Phylogeny of Y-chromosome SNPs and haplogroup frequencies in different Italian populations.

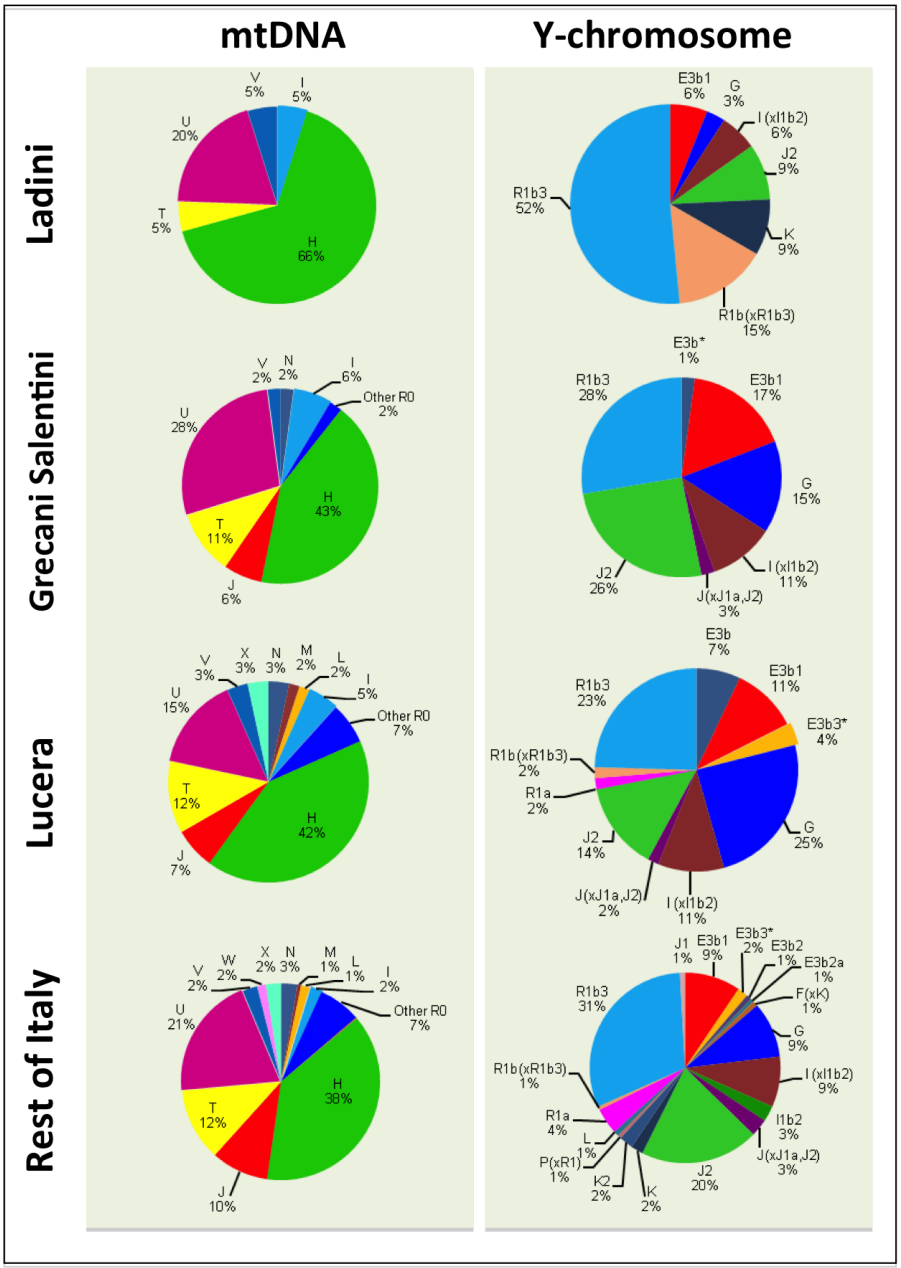


Figure 4. Haplogroup frequencies of Ladin, Grecani Salentini and Lucera compared to the rest of the Italian populations analyzed in the present study.

Legend to Supplementary Data

Table S1. References to the populations samples used in the present study for population comparison analysis.

Table S2. Mitochondrial DNA control region haplotypes obtained in the samples analyzed in the present study.

Table S3. mtSNPs and primers used to characterize J/T and U and some of their sub-clades.

PROCEEDINGS OF THE ROYAL SOCIETY B

FirstCite e-publishing

Proc. R. Soc. B
doi:10.1098/rspb.2011.1044
Published online

The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269

George B. J. Busby[1], Francesca Brisighelli[1,3,4], Paula Sánchez-Diz[3], Eva Ramos-Luis[3], Conrado Martinez-Cadenas[1], Mark G. Thomas[6], Daniel G. Bradley[7], Leonor Gusmão[8], Bruce Winney[2], Walter Bodmer[2], Marielle Vennemann[9,10], Valentina Coia[4,11], Francesca Scarnicci[12], Sergio Tofanelli[13], Giuseppe Vona[14], Rafal Ploski[15], Carla Vecchiotti[5], Tatijana Zemunik[16], Igor Rudan[16,17], Sena Karachanak[18], Draga Toncheva[18], Paolo Anagnostou[4,19], Gianmarco Ferri[20], Cesare Rapone[21], Tor Hervig[22], Torolf Moen[23], James F. Wilson[17,24] and Cristian Capelli[1,*]

[1]*Department of Zoology, and* [2]*Department of Clinical Pharmacology, University of Oxford, Oxford, UK*
[3]*Institute of Forensic Sciences Luis Concheiro, Genomics Medicine Group, University of Santiago de Compostela, Spain*
[4]*Department of Environmental Biology, and* [5]*Department of Anatomy, Histology, Legal Medicine and Locomotory System, University of Rome 'La Sapienza', Rome, Italy*
[6]*Department of Genetics, Evolution and Environment, University College London, London, UK*
[7]*Smurfit Institute of Genetics, Trinity College Dublin, Republic of Ireland*
[8]*IPATIMUP, Institute of Pathology and Molecular Immunology of the University of Porto, Portugal*
[9]*Centre for Forensic Science, University of Strathclyde, Glasgow, UK*
[10]*Institute of Legal Medicine, University of Freiburg, Freiburg, Germany*
[11]*Department of Philosophy, History and Cultural heritage, University of Trento, Trento, Italy*
[12]*Department of Legal Medicine, University 'Cattolica del Sacro Cuore', Rome, Italy*
[13]*Department of Biology, University of Pisa, Pisa, Italy*
[14]*Department of Experimental Biology, University of Cagliari, Monserrato-Cagliari, Italy*
[15]*Department of Medical Genetics, Warsaw Medical University, Warsaw, Poland*
[16]*Croatian Centre for Global Health, University of Split School of Medicine, Split, Croatia*
[17]*Centre for Population Health Sciences, University of Edinburgh, Teviot Place, Edinburgh, Scotland, UK*
[18]*Department of Medical Genetics, Medical University of Sofia, Sofia, Bulgaria*
[19]*Department of Evolutionary and Experimental Biology, University of Bologna, Italy*
[20]*Section of Legal Medicine, Department of Diagnostic and Laboratory Services and Legal Medicine, University of Modena and Reggio Emilia, Modena, Italy*
[21]*Biology Section, Carabinieri Scientific Research Department, Rome, Italy*
[22]*The Gade Institute, University of Bergen, Norway*
[23]*Department of Laboratory Medicine, Children's and Women's Health, Faculty of Medicine, NTNU, Trondheim, Norway*
[24]*Ethnoancestry Limited, Edinburgh, UK*

Recently, the debate on the origins of the major European Y chromosome haplogroup R1b1b2-M269 has reignited, and opinion has moved away from Palaeolithic origins to the notion of a younger Neolithic spread of these chromosomes from the Near East. Here, we address this debate by investigating frequency patterns and diversity in the largest collection of R1b1b2-M269 chromosomes yet assembled. Our analysis reveals no geographical trends in diversity, in contradiction to expectation under the Neolithic hypothesis, and suggests an alternative explanation for the apparent cline in diversity recently described. We further investigate the young, STR-based time to the most recent common ancestor estimates proposed so far for R-M269-related lineages and find evidence for an appreciable effect of microsatellite choice on age estimates. As a consequence, the existing data and tools are insufficient to make credible estimates for the age of this haplogroup, and conclusions about the timing of its origin and dispersal should be viewed with a large degree of caution.

Keywords: Y-STRs; R1b1b2-M269; neolithic hypothesis; average squared distance

*Author for correspondence (cristian.capelli@zoo.ox.ac.uk).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2011.1044 or via http://rspb.royalsocietypublishing.org.

Received 17 May 2011
Accepted 3 August 2011

1. INTRODUCTION

Since the first attempts to use biological variation in humans to aid our understanding of early human migrations, the peopling of Europe has been a major research focus [1,2]. Following the development of agriculture in the Fertile Crescent some 10 000 years ago [3,4], this technology spread from the Near East westward into Europe, causing a major cultural transition from itinerant hunter–gathering to sedentary farming, which led to dramatic population growth [5,6], during what has become known as the Neolithic transition [7,8]. Within this archaeological framework, debate rages about the relative contributions to modern European populations of the first people of Europe and those who migrated into it with the Neolithic transition, both in terms of their genetic legacy and as to the processes of migration and succession [9–16]. The true scenario is undoubtedly multi-faceted and complex. Both early work on 'classical markers' using principal components analysis and more recent studies using the Y chromosome have shown that in Europe, genetic variation is distributed along a southeast–northwest gradient. Such observations have been suggested to support a model of demic diffusion for the Neolithic transition in Europe (i.e. that the spread of agriculture also involved an associated movement of people from the Near East) [2,17–19].

New work [20–22] has addressed the Neolithic transition in Europe by focusing on the main western European Y chromosome haplogroup R1b1b2-M269 (hereafter referred to as R-M269). This lineage had hitherto received little recent attention in this context, although previous work suggested that the broader R-M173 clade (excluding the R1a-M17 sub-lineage) and Haplogroup 1 (derived at single nucleotide polymorphism, or SNP, 92r7) are likely to have spread into Europe during the Palaeolithic [17,18,23], and therefore unlikely to have been carried into Europe with the migrating farmers. Balaresque *et al.* [20] (hereafter 'Balaresque') used 840 Y chromosomes within haplogroup R-M269 to show that, although this haplogroup is characterized by a strong frequency cline from high in the west to low in the east, the associated cline in haplotype diversity (measured as mean short tandem repeat, or STR, variance) is in the opposite direction. They posited that this correlation could be explained by a more recent dispersal of this lineage from the Near East coinciding with the Neolithic transition in Europe. The lineage was estimated to be approximately 6000 years old in various populations, which was argued to be consistent with this model. This result, as noted in their introduction, 'indicates that the great majority of the Y chromosomes of Europeans have their origins in the Neolithic expansion' (p. 2 in [20]).

Myres *et al.* [21] described several new SNP mutations downstream of R-M269 that show strong geographical structuring in a much larger sample of 2043 R-M269 chromosomes. They highlight an essentially European-specific clade, defined by the presence of SNPs M412 (also known as S167) and L11 (S127), which is clinal from high frequencies (greater than 70%) in western Europe, decreasing eastward. This study showed that the distributions of several downstream SNPs exhibit striking frequency patterns and appear to spread from different areas of highly localized frequencies, some of which were also observed by Cruciani *et al.* [24]. Myres *et al*. estimated coalescence times for the R-S116 haplogroup in different populations in Europe and suggested, in broad agreement with Balaresque, that the R-M269 haplogroup may have spread with the Neolithic, and more specifically with the *Linearbandkeramik*, a Neolithic agricultural industry that spread throughout northern Europe, from Hungary to France, around 7500 years ago.

The current uncertainty surrounding STR mutation rates shows that despite these recent studies, there can still be no consensus on when and where the R-M269 haplogroup originated and spread in Europe. Even if invoking the origins of the European Y chromosome gene pool 'must be viewed cautiously especially when such an argument is based on just a single incompletely resolved haplogroup' (p. 100 in [21]), it is of profound interest to try to understand how the vast majority of western European men (greater than 100 million) carry Y chromosomes that belong to the R-M269 Y chromosome haplogroup.

Consequently, we have addressed these issues with our own large R-M269 dataset, both on its own and in combination with compatible data from the most recent comprehensive survey [21]. We show that the fundamental relationship between mean STR variance and longitude, which is the basis of the recent claim of support for the Neolithic hypothesis [20], does not hold for our larger and geographically broader sample. We also explain how this previous analysis may have resulted in this spurious association. We finally explore the spatial distribution of genetic diversity associated with the R-M269 European-specific sub-lineage, defined by SNP S127, showing an essentially homogeneous background of microsatellite variation at several different sub-lineage levels, based on a common set of 10 STRs typed across 2000 R-M269 chromosomes.

While acknowledging uncertainty, researchers usually report the age of Y chromosome lineages based on differences between individuals across multiple STRs, often using average squared distance (ASD) or related summary statistics [25,26] as unbiased estimators of coalescence time, *T*. We investigated how ASD changes in our dataset based on different sets of STRs. Contrary to common belief, estimates of ASD, and therefore *T*, vary widely when different subsets of STRs are used with the same sample. While recent evidence has increased support for the Neolithic spread of R-M269, we conclude that at the present time it is not possible to make any credible estimate of divergence time based on the sets of Y-STRs used in recent studies. Furthermore, we show that it is the properties of Y-STRs, not the number used *per se*, that appear to control the accuracy of divergence time estimates, attributes which are rarely, if ever, considered in practise.

2. MATERIAL AND METHODS

(a) *Ethics statement*

All males sampled gave informed consent following ethical approval by the ethics committees at the various universities where the samples were collected.

(b) *DNA samples and genotyping*

We assembled a dataset of 2486 R-M269 Y chromosomes from across Europe, the Near East and western Asia, from a total population of 6503, which included both novel and previously published Y chromosomes. To assess the frequency

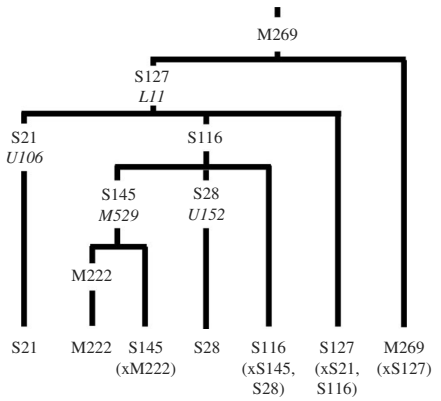


Figure 1. Y chromosome tree showing the relationships of SNPs downstream from R-M269 tested in this study. Alternative nomenclature for some SNPs is provided in italics.

distribution of R-M269 and various sub-haplogroups in Europe and Asia, we combined our data with that of Myres *et al.* [21], which gave a combined set of 4529 R-M269 chromosomes from a total sample of 16 298 from 172 different populations (electronic supplementary material, table S1 and figure S1). The frequencies of the following SNPs, whose phylogeny is shown in figure 1, were ascertained: S127/L11 (rs9786076), S21/U106 (rs16981293), S116 (rs34276300), S145/M529 (rs11799226) and S28/U152 (rs1236440). Samples were amplified in a standard PCR reaction and the SNaPshot Multiplex System (Life Technologies Corp., Carlsbad, CA, USA) primer extension protocol was used to characterize the allele present at each SNP loci. All primers are listed in the electronic supplementary material.

For the majority of the individuals typed in this study (2289), the following 10 STRs were available: DYS19; DYS389I; DYS389b (subtracting the alleles scored at DYS389I from the DYS389II locus); DYS390; DYS391; DYS392; DYS393; DYS437; DYS438; and DYS439, either being previously published or having been typed by ourselves using the Yfiler kit (Life Technologies Corp.) [27] or the Promega Powerplex assay (Promega Corp., Madison, WI, USA) [28]. For the samples from Weale *et al.* [29], only five STRs were previously published, and so the remaining five were typed with an internally designed and verified multiplex using primers from the study of Butler *et al.* [30] for DYS391, DYS437, DYS389I and II and DYS439, and primers from the study of Gusmao & Alves [31] for DYS438. DYS391 calls were used to check for consistency with the original haplotypes of Weale *et al.* Three of the Weale *et al.* populations were not typed further for these STRs (114 individuals). Individuals typed using the Yfiler kit (1035) were used to investigate the effect of STR selection on ASD calculations (electronic supplementary material, table S2).

Populations with a total size of 30 or above were used to build the frequency maps (electronic supplementary material, figure S1). Variance was calculated only for those populations where haplotypes were available for at least 10 individuals within the relevant haplogroup.

(c) *Analysis*

Maps of SNP frequencies were displayed using ArcMap GIS (v. 9.2; ESRI). Interpolation was performed using the inverse distance weighting procedure. Latitudes and longitudes for all populations were based on the highest-resolution sampling centre associated with the samples and are shown in electronic supplementary material, table S1.

The R statistical package [32] was used to calculate the median STR variance (the variance in the number of repeats within a locus averaged across all loci) between all individuals within a population following 1000 bootstrap replicates with replacement over individuals. Regression analysis was performed in R to compare average STR variance with latitude and longitude for the R-M269, R-M269(xS127) and R-S127 haplogroups.

We investigated how ASD estimates change within our sample when using different combinations of STRs based on two separate criteria: mutation rate, μ; and observed linearity, $\theta(R)$ (table 1). We used the observed μ calculated recently [33] to rank the 15 STRs on a scale of speed, and separately calculated ASD based on the seven fastest and seven slowest rates (electronic supplementary material, table S4). Our second criterion was based on the estimated duration of linearity, D, of different groups of STRs. Duration of linearity is an estimate of the divergence time after which ASD ceases to increase linearly with time. For STRs mutating under a strict stepwise model, Goldstein *et al.* showed that ASD initially increases linearly with time, but that this linearity is constrained by the maximum number of repeats an STR can take, R [26]. D is approximated using $\theta(R)$ (which is a simple transformation of R) and μ, and the effective population size (Ne) (eqns 3 and 4 in [26]). Greater values of $\theta(R)/2\mu$ yield increased estimates of D. Using STRs with greater values of $\theta(R)/2\mu$ should allow linearity to be assumed further into the past, and ASD calculated from these STRs should be less likely to be underestimated as a result of saturation. Table 1 and electronic supplementary material, table S4 show the different groups of STRs used and associated values of μ, R, $\theta(R)/2\mu$ and ASD.

To check that any differences in time to the most recent common ancestor (TMRCA) estimation are not specific to methods based on ASD, we used BATWING [35] on the HGDP Bedouin population for which a greater number of Y-STRs ($n = 65$) were available [36]. We compared four different sets of STRs with varying degrees of duration of linearity estimates (electronic supplementary material).

3. RESULTS

To investigate the origins of the R-M269 lineage in Europe, we analysed a large dataset of 4529 R-M269 chromosomes (2486 of which have not previously been published at such detailed resolution) from several populations across Europe, the Near East and western Asia (electronic supplementary material, figure S1 and table S1). Within Europe, we observed a northwest–southeast frequency cline for R-M269, similar to those observed previously [10,11,37], from high frequencies in western Europe to lower frequencies in the east. Within haplogroup R-M269 we genotyped a newly characterized SNP, S127 (equivalent to L11), for which the distribution in Europe and the Near East, together with that of R-M269 and R-M269(xS127), are shown in figure 2. The distributions of R-M269 and R-S127 are broadly

Table 1. Fifteen Y-STRs with mutation rates, range of alleles and estimate of duration of linearity. All STRs investigated in this study are shown with their mutation rates (μ), estimated from Ballantyne *et al.* [33], and range of observed alleles, R, with 95% CI is taken from the YHRD [34]. $\theta(R)/2\mu$ is an estimate of the duration of linearity of an STR (see §2).

Y-STR	μ	$\mu(2.5)$	$\mu(97.5)$	R	$\theta(R)/2\mu$
DYS448	0.000394	0.0000141	0.00211	11	25 381
DYS392	0.00097	0.000143	0.00323	15	19 244
DYS438	0.000956	0.000137	0.00318	12	12 465
DYS390	0.00152	0.000352	0.00409	13	9211
DYS393	0.00211	0.000621	0.005	12	5648
DYS439	0.00384	0.00163	0.00754	15	4861
DYS437	0.00153	0.000354	0.0041	9	4357
DYS635	0.00385	0.00163	0.00755	14	4221
DYS456	0.00494	0.00235	0.00897	14	3289
DYS389II	0.00383	0.00161	0.00749	12	3111
DYS391	0.00323	0.00126	0.00665	10	2554
DYS458	0.00836	0.0048	0.0134	14	1944
DYS19	0.00437	0.00198	0.00823	10	1888
Y-GATA-H4	0.00322	0.00128	0.00662	8	1630
DYS389I	0.00551	0.00272	0.00974	8	953

overlapping, but the frequency of R-S127 drops off around the Balkans, reaching extremely low values further to the east and outside of Europe. Conversely, R-M269(xS127) shows higher frequencies in eastern populations. Frequency maps showing three geographically localized R-S127 sub-haplogroups (R-S21, R-S145 and R-S28) are shown in figure 3.

We next calculated STR diversity for each population for the whole R-M269 lineage, and for the R-S127 and R-M269(xS127) sub-haplogroups, and investigated the relationship between average STR variance and longitude and latitude in exactly the same fashion as Balaresque. We provide estimates of uncertainty for these values by bootstrapping over individuals, and report the median of the observed variance values and its 95 per cent CI (figure 2). We normalized latitude and longitude, and performed a linear regression between these values and the median microsatellite variance for the three R-M269 sub-haplogroups. We found no correlation with latitude (data not shown) and, contrary to Balaresque, we did not find any significant correlation between longitude and variance for any haplogroup.

The Balaresque dataset presents genotype data only to the resolution of SNP R-M269. Our results show that the vast majority of R-M269 samples in Anatolia, approximately 90 per cent, belong to the R-M269(xS127) sub-haplogroup. Removing these Turkish populations from the Balaresque data and repeating the regression removes the significant correlation ($R^2 = 0.23$, $p = 0.09$; details in the electronic supplementary material and figure S2). These populations are therefore intrinsic to the significant correlation.

We observed that the Irish haplotypes used in the Balaresque analysis had a very low STR variance (0.208) compared with those included in our analysis (0.35; originally published by Moore *et al.* [38]). Balaresque used a sample of Irish haplotypes downloaded from the online Ysearch database (http://www.ysearch.org). To test if the Ysearch haplotypes were representative of the Irish R-M269 of Moore *et al.* [38], we independently resampled the Moore *et al.* dataset 10 000 times, selecting sub-samples of 75 haplotypes from which we estimated the variance using the same nine STRs used in the Balaresque paper (detailed methodology and justification can be found in the electronic supplementary material). The median variance of these 10 000 repetitions was 0.354 with a 95 per cent CI of (0.285–0.432). When we repeated the regression analysis with this different variance estimate, the correlation was no longer significant ($R^2 = 0.09$, $p = 0.19$).

Microsatellite-based ASD has been shown to increase linearly with time [26] and has been used as an unbiased estimator of mean coalescence time, given that it approximates to $2\mu T$ [21,25,39]. It would be expected that using different sets of STRs should not dramatically alter the estimation of T: as μ changes, ASD should similarly change, with T staying constant. Table 1 shows estimates of the duration of linearity based on observed mutation rates estimated recently [33] and range estimated from the YHRD [34]. The ASD for R-S127 was calculated by comparing the 15 STR haplotypes of its two major sub-haplogroups, R-S21 (141 chromosomes) and R-S116 (717; electronic supplementary material, table S3). Figure 4*a* is a plot of T (estimated as ASD/2μ) for several different sets of STRs with different characteristics (electronic supplementary material, table S4).

To further explore the correlation between T and STR selection, we calculated T in the same way as described above based on chromosomes belonging to the two deepest branches of the Y chromosome phylogeny, AxA1 and B [40] (figure 4*b*; electronic supplementary material, table S4). As a comparison, ASD calculated from the same STR subsets is shown for the R-S127 on the same plot.

4. DISCUSSION

Here, we have confirmed with the broadest analysis to date that the spatial distribution of Y chromosome haplogroup M269 can be split by R-S127 into European and western Eurasian lineages. Contrary to the results of Balaresque, we see no relationship between diversity

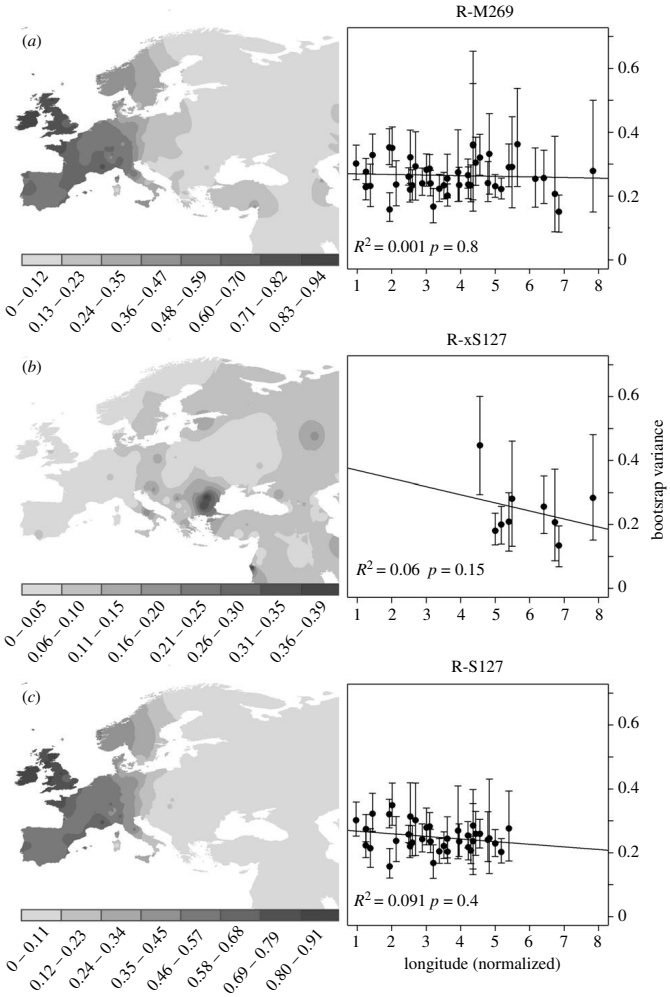


Figure 2. Frequency distributions and variation of Y chromosome haplogroups R-M269, R-S127 and R-M269(xS127) in Europe. The three panels show contour maps based on the frequencies of the different haplogroups found across Europe and western Asia: (*a*) R-M269, (*b*) R-S127 and (*c*) R-M269(xS127). The maps on the left are based on the frequencies of the SNPs in all populations marked on the map (data in electronic supplementary material, table S1 and figure S1). The graphs on the right show the relationship between longitude and bootstrap variance based on 10 STRs for all populations with at least 10 individuals carrying that SNP. The R^2 and associated *p*-values are shown for the correlations in the graphs. The population codes are detailed in table 1 and electronic supplementary material, table S1.

and longitude (figure 2) for R-M269. The presence of two sets of populations in the Balaresque paper appears to be causal to the observed relationship: the underestimated diversity of the Irish population and the inclusion of the Turkish chromosomes, the majority of which potentially belong to the non-European clade R-M269(xS127). When these elements are properly taken into account, jointly or independently, the correlation no longer exists. This correlation is the central tenet to the hypothesis that R-M269 was spread with expanding Neolithic farmers.

Morelli *et al.* [22] (hereafter 'Morelli') found STR motifs that split R-M269 into eastern and western lineages. We observed that 71 per cent of the Myres *et al.* R-M269(xS127) chromosomes for which STR information is available have the eastern motif (DYS393-12/DYS461-10), while 80 per cent of the R-S127 chromosomes of Myres *et al.* have the western

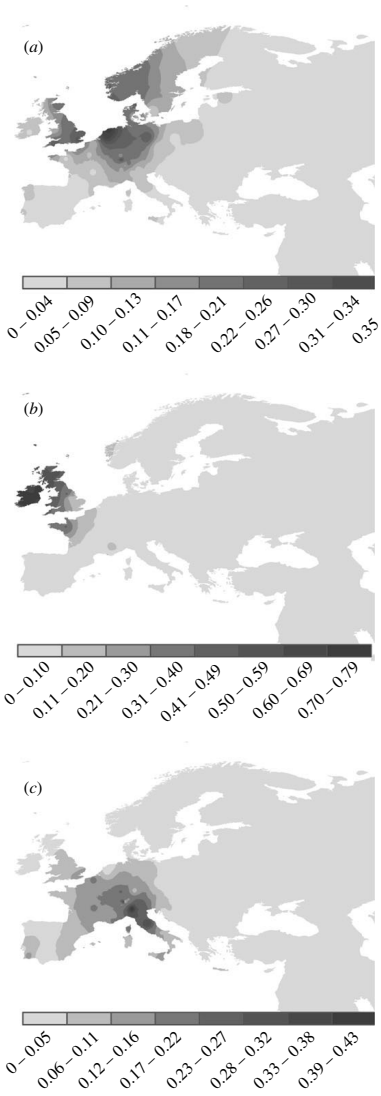


Figure 3. Frequency distributions of R-M269 sub-haplogroups. Contour maps for lineages defined by marker (*a*) R-S21, (*b*) R-S145 and (*c*) R-S28.

motif (DYS393-13/DYS461-11). No R-S127 chromosomes displayed the eastern motif, while 5 per cent of R-M269(xS127) chromosomes displayed the western motif (all of which were either L23 (S141) or M412 (S127)-derived). In both cases, however, these motifs differed from those suggested by Morelli by having one repeat less at the DYS461 locus. The dichotomy observed by Morelli based on a two STR motif is therefore corroborated, at least in part, by the presence of this SNP.

Dating of Y chromosome lineages is notoriously controversial [25,41–44], the major issue being that the choice of STR mutation rate can lead to age estimates that differ by a factor of three (i.e. the evolutionary [25] versus observed (genealogical) mutation rates [33,45]). Interestingly, despite the fact that Myres *et al.* and Balaresque used different STR mutation rates and dating approaches, their TMRCA estimates overlap: 8590–11 950 years using a mutation rate of 6.9×10^{-4} per generation, and 4577–9063 years using an average mutation rate of 2.3×10^{-3}, respectively. Separately, Morelli calculated the TMRCA based only on Sardinian and Anatolian chromosomes, and estimated the R-M269 lineage to have originated 25 000–80 700 years ago) [22], based on the same evolutionary mutation rate [25,41] as Myres *et al.*

In seeking to find a suitable set of STRs with which to estimate the average coalescence time, *T*, of sub-haplogroup R-S127, we have shown that not all STRs are of equal use in this context. We concentrated on estimating the duration of linearity, *D*, using different sets of STRs. Our analyses suggest that the *D* of an STR is key to its ability to uncover deep ancestry. Duration of linearity refers to the length of time into the past over which ASD and *T* continue to be linearly related for a specific STR. Goldstein *et al.* [26] showed that *D* is affected by two properties of the STRs used to calculate ASD: the mutation rate and range of possible alleles that the STR can take. When we manipulated our choice of STR marker based on $\theta(R)/2\mu$ (a surrogate for *D*; table 1), we found that different sets of STRs gave different values for *T*. It is clear, then, that coalescence estimates explicitly depend on the STRs that one uses.

Our analysis confirms that this phenomenon is not specific to the R-M269 haplogroup nor to methods using ASD. Figure 4*b* shows that STRs with high *D* produce larger estimates of *T*. What is clear is that estimates of *T* implicitly depend on the STRs that are selected to make this inference. Using BATWING on an HGDP population for which 65 Y-STRs are available, we have shown that the median estimate of TMRCA can differ by over five times when STRs are selected on the basis of the expected duration of linearity (electronic supplementary material, figure S4). While researchers take into account STR mutation rates when estimating divergence time with ASD, commonly used STRs do not have the specific attributes that allow linearity to be assumed further into the past. The majority of haplogroup dates based on such sets of STRs may therefore have been systematically underestimated.

5. CONCLUSION

The distributions of the main R-S127 sub-haplogroups, R-S21, R-S145 and R-S28, show markedly localized concentrations (figure 3). If the R-M269 lineage is more recent in origin than the Neolithic expansion, then its current distribution would have to be the result of major population movements occurring since that origin. For this haplogroup to be so ubiquitous, the population carrying R-S127 would have displaced most of the populations present in western Europe after the Neolithic agricultural transition. Alternatively, if R-S127 originated prior to the Neolithic wave of

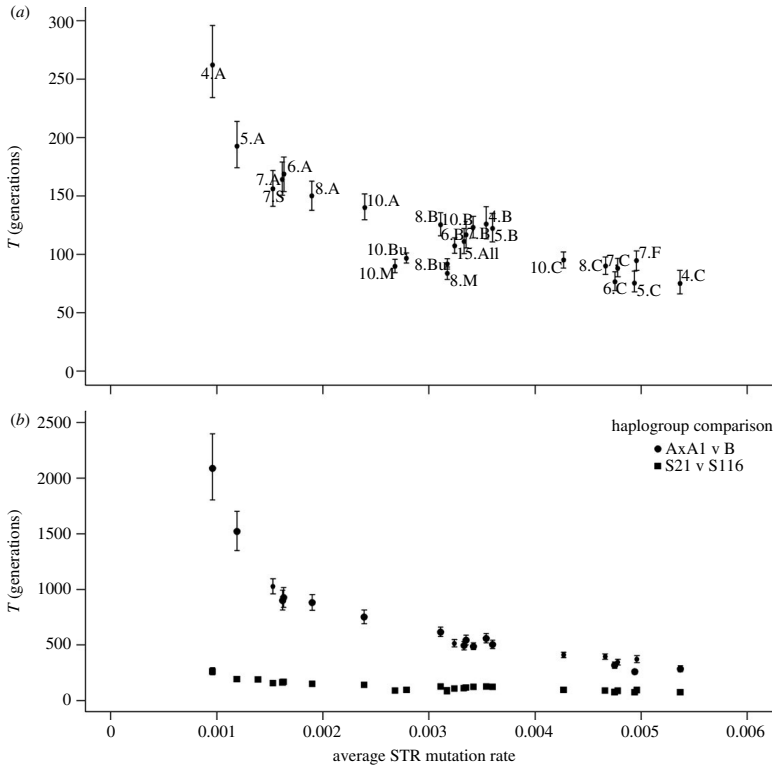


Figure 4. Relationship between time to the most recent common ancestor, *T*, and mutation rate, μ, for various STR subsets. (*a*) Estimates of *T* for the R-S127 haplogroup. Points are labelled with the subset of STRs used to calculate *T* and are detailed in electronic supplementary material, table S4. (*b*) The same data, but this time together with estimates of *T* based on comparisons of Y chromosome A and B haplogroups (see main text).

expansion, then either it was already present in most of Europe before the expansion, or the mutation occurred in the east, and was spread before or after the expansion, in which case we would expect higher diversity in the east closer to the origins of agriculture, which is not what we observe. The maps of R-S127 sub-haplogroup frequencies for R-S21, R-S145 and R-S28 show radial distributions from specific European locations (figure 3). These centres have high absolute frequencies: R-S21 has a frequency of 44 per cent in Friesland, and R-S28 reaches 25 per cent in the Alps; and in the populations where they are at the highest frequency, the vast majority of R-S127 belong to that particular sub-lineage. For example, half of all R-M269 across southern Europe is R-S28-derived, and around 60 per cent of R-M269 in Central Europe is R-S21-derived. At the sub-haplogroup level, then, R-M269 is split into geographically localized pockets with individual R-M269 sub-haplogroups dominating, suggesting that the frequency of R-M269 across Europe could be related to the growth of multiple, geographically specific sub-lineages that differ in different parts of Europe.

A recent analysis of radiocarbon dates of Neolithic sites across Europe [46] reveals that the spread of the Neolithic was by no means constant, and that several 'centres of renewed expansion' are visible across Europe, representing areas of colonization, three of which map intriguingly closely to the centres of the sub-haplogroups foci (electronic supplementary material, figure S3). Future work involving spatially explicit simulations, together with accurate measures of Y chromosome diversity, are needed to investigate how the current distribution of sub-haplogroups may have been produced. In this context, recent work by Sjödin & François [47] rejected a Palaeolithic dispersion for R1b-M269 using spatial simulations based on the dataset of Balaresque. Nevertheless, we note that additional work is still necessary as these authors were not aware of the limitation of the Balaresque dataset presented here, and did not fully explore the impact of the different molecular characteristics of the investigated loci on their analysis.

Age estimates based on sets of Y-STRs carefully selected to possess the attributes necessary for uncovering deep ancestry (for example, from the almost 200 recently characterized here [33]), and from whole Y chromosome

sequence comparisons, will provide robust dates for this haplogroup in the future. For now, we can offer no date as to the age of R-M269 or R-S127, but believe that our STR analyses suggest the recent age estimates of R-M269 [20] and R-S116 [21] are likely to be younger than the true values, and the homogeneity of STR variance and distribution of sub-types across the continent are inconsistent with the hypothesis of the Neolithic diffusion of the R-M269 Y chromosome lineage.

We thank all donors who contributed DNA for this project and Hugh Sturrock for help with the spatial analysis. We also thank Prof. Bernd Brinkmann at the Institute of Forensic Genetics in Münster for access to samples from Turkey. G.B.J.B. and C.C. conceived and designed the experiments, G.B.J.B., F.B., P.S.D., E.R.L., C.M.C., M.G.T., D.G.B., L.G., M.V., G.F. and J.F.W. genotyped the samples, G.B.J.B. analysed the data, C.M.C., M.G.T., D.G.B., L.G., B.W., W.B., M.V., V.C., F.S., S.T., G.V., R.P., C.V., T.Z., I.R., S.K., D.T., P.A., G.F., C.R., T.H., T.M. and J.F.W. contributed reagents/materials, and G.B.J.B. and C.C. wrote the paper.

G.B.J.B. is supported by a BBSRC doctoral training grant and Somerville College, University of Oxford, and J.F.W. by the Royal Society. A subset of the genotyping data was generated within a project funded by the British Academy (BARDA-47870). V.C. was supported by Provincia Autonoma di Trento (BIOSTRE project, Post-doc 2006). C.C. is an RCUK Academic Fellow. F.B., V.C. and P.A. were supported by grant Prin 2009 project 200975T9EW from MUIR. S.T. was supported by a University of Pisa 60%2010 grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare no conflict of interest.

REFERENCES

1 Menozzi, P., Piazza, A. & Cavalli-Sforza, L. 1978 Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792. (doi:10.1126/science.356262)
2 Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. 1994 *The history and geography of human genes*. Princeton, NJ: Princeton University Press.
3 Diamond, J. & Bellwood, P. 2003 Farmers and their languages: the first expansions. *Science* **300**, 597–603. (doi:10.1126/science.1078208)
4 Blockley, S. P. E. & Pinhasi, R. 2011 A revised chronology for the adoption of agriculture in the Southern Levant and the role of Late Glacial climatic change. *Q. Sci. Rev.* **30**, 98–108. (doi:10.1016/j.quascirev.2010.09.021)
5 Gamble, C., Davies, W., Pettitt, P., Hazelwood, L. & Richards, M. 2005 The archaeological and genetic foundations of the European population during the Late Glacial: implications for 'agricultural thinking'. *Camb. Archaeol. J.* **15**, 193–223. (doi:10.1017/S0959774305000107)
6 Collard, M., Edinborough, K., Shennan, S. & Thomas, M. G. 2010 Radiocarbon evidence indicates that migrants introduced farming to Britain. *J. Archaeol. Sci.* **37**, 866–870. (doi:10.1016/j.jas.2009.11.016)
7 Cunliffe, B. 1994 *The Oxford illustrated history of prehistoric Europe*. Oxford, UK: Oxford University Press.
8 Jobling, D. M., Hurles, M. & Tyler-Smith, C. 2004 *Human evolutionary genetics: origins, peoples and disease*, 1st edn. New York, NY: Garland Science.
9 Chikhi, L., Nichols, R. A., Barbujani, G. & Beaumont, M. A. 2002 Y genetic data support the Neolithic demic diffusion model. *Proc. Natl Acad. Sci. USA* **99**, 11 008–11 013. (doi:10.1073/pnas.162158799)
10 Capelli, C. *et al.* 2003 A Y chromosome census of the British Isles. *Curr. Biol.* **13**, 979–984. (doi:10.1016/S0960-9822(03)00373-7)
11 Capelli, C. *et al.* 2006 Population structure in the Mediterranean basin: a Y chromosome perspective. *Ann. Hum. Genet.* **70**, 207–225. (doi:10.1111/j.1529-8817.2005.00224.x)
12 Battaglia, V. *et al.* 2008 Y-chromosomal evidence of the cultural diffusion of agriculture in southeast Europe. *Eur. J. Hum. Genet.* **17**, 820–830. (doi:10.1038/ejhg.2008.249)
13 Gallagher, A., Gunther, M. M. & Bruchhaus, H. 2009 Population continuity, demic diffusion and Neolithic origins in central-southern Germany: the evidence from body proportions. *HOMO J. Comp. Hum. Biol.* **60**, 95–126. (doi:10.1016/j.jchb.2008.05.006)
14 Francalacci, P. & Sanna, D. 2008 History and geography of human Y-chromosome in Europe: a SNP perspective. *J. Anthropol. Sci.* **86**, 59–89.
15 Rowley-Conwy, P. 2009 Human prehistory: hunting for the earliest farmers. *Curr. Biol.* **19**, R948–R949. (doi:10.1016/j.cub.2009.09.054)
16 Francalacci, P., Morelli, L., Useli, A. & Sanna, D. 2010 The history and geography of the y chromosome SNPs in Europe: an update. *J. Anthropol. Sci.* **88**, 207–214.
17 Semino, O. *et al.* 2000 The genetic legacy of paleolithic Homo sapiens sapiens in extant europeans: a Y chromosome perspective. *Science* **290**, 1155–1159. (doi:10.1126/science.290.5494.1155)
18 Rosser, Z. H. *et al.* 2000 Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.* **67**, 1526–1543. (doi:10.1086/316890)
19 Novembre, J. *et al.* 2008 Genes mirror geography within Europe. *Nature* **456**, 98–101. (doi:10.1038/nature07331)
20 Balaresque, P. *et al.* 2010 A predominantly neolithic origin for European paternal lineages. *PLoS Biol.* **8**, e1000285. (doi:10.1371/journal.pbio.1000285)
21 Myres, N. M. *et al.* 2011 A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur. J. Hum. Genet.* **19**, 95–101. (doi:10.1038/ejhg.2010.146)
22 Morelli, L., Contu, D., Santoni, F., Whalen, M. B., Francalacci, P. & Cucca, F. 2010 A comparison of Y-chromosome variation in Sardinia and Anatolia is more consistent with cultural rather than demic diffusion of agriculture. *PLoS ONE* **5**, e10419. (doi:10.1371/journal.pone.0010419)
23 Wilson, J. F., Weiss, D. A., Richards, M., Thomas, M. G., Bradman, N. & Goldstein, D. B. 2001 Genetic evidence for different male and female roles during cultural transitions in the British Isles. *Proc. Natl Acad. Sci. USA* **98**, 5078–5083. (doi:10.1073/pnas.071036898)
24 Cruciani, F. *et al.* 2011 Strong intra- and inter-continental differentiation revealed by Y chromosome SNPs M269, U106 and U152. *Forensic Sci. Int. Genet.* **5**, e49–e52. (doi:10.1016/j.fsigen.2010.07.006)
25 Zhivotovsky, L. A. *et al.* 2004 The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* **74**, 50–61. (doi:10.1086/380911)
26 Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. 1995 An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463.
27 Mulero, J. J., Chang, C. W., Calandro, L. M., Green, R. L., Li, Y., Johnson, C. L. & Hennessy, L. K. 2006 Development and validation of the AmpF*l*STR® Yfiler™ PCR amplification kit: a male specific, single amplification 17

Y-STR multiplex system. *J. Forensic Sci.* **51**, 64–75. (doi:10.1111/j.1556-4029.2005.00016.x)
28 Krenke, B. E. *et al.* 2005 Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. *Forensic Sci. Int.* **148**, 1–14. (doi:10.1016/j.forsciint.2004.07.008)
29 Weale, M. E., Weiss, D. A., Jager, R. F., Bradman, N. & Thomas, M. G. 2002 Y chromosome evidence for Anglo-Saxon mass migration. *Mol. Biol. Evol.* **19**, 1008–1021.
30 Butler, J. M., Schoske, R., Vallone, P. M., Kline, M. C., Redd, A. J. & Hammer, M. F. 2002 A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. *Forensic Sci. Int.* **129**, 10–24. (doi:10.1016/S0379-0738(02)00195-0)
31 Gusmao, L. & Alves, C. 2005 Y Chromosome STR typing. In *Forensic DNA typing protocols* (ed. A. Carracedo), pp. 67–81. Totowa, NJ: Humana Press.
32 R Development Core Team. 2011 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
33 Ballantyne, K. N. *et al.* 2010 Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.* **87**, 341–353. (doi:10.1016/j.ajhg.2010.08.006)
34 Willuweit, S. & Roewer, L. 2007 Y chromosome haplotype reference database (YHRD): update. *Forensic Sci. Int. Genet.* **1**, 83–87. (doi:10.1016/j.fsigen.2007.01.017)
35 Wilson, I. J., Weale, M. E. & Balding, D. J. 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Stat. Soc. A* **166**, 155–188. (doi:10.1111/1467-985X.00264)
36 Shi, W. *et al.* 2010 A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Mol. Biol. Evol.* **27**, 385–393. (doi:10.1093/molbev/msp243)
37 Capelli, C. *et al.* 2007 Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic–Neolithic encounter. *Mol. Phylogenet. Evol.* **44**, 228–239. (doi:10.1016/j.ympev.2006.11.030)
38 Moore, L. T., McEvoy, B., Cape, E., Simms, K. & Bradley, D. G. 2006 A Y-chromosome signature of hegemony in Gaelic Ireland. *Am. J. Hum. Genet.* **78**, 334–338. (doi:10.1086/500055)
39 Sengupta, S. *et al.* 2006 Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am. J. Hum. Genet.* **78**, 202–221. (doi:10.1086/499411)
40 Batini, C. *et al.* In press. Signatures of the pre-agricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Mol. Biol. Evol.* (doi:10.1093/molbev/msr089)
41 Zhivotovsky, L. A., Underhill, P. A. & Feldman, M. W. 2006 Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. *Mol. Biol. Evol.* **23**, 2268–2270. (doi:10.1093/molbev/msl105)
42 Di Giacomo, F. *et al.* 2004 Y chromosomal haplogroup J as a signature of the post-Neolithic colonization of Europe. *Hum. Genet.* **115**, 357–371. (doi:10.1007/s00439-004-1168-9)
43 Zhivotovsky, L. A. & Underhill, P. A. 2005 On the evolutionary mutation rate at Y-chromosome STRs: comments on paper by Di Giacomo et al. (2004). *Hum. Genet.* **116**, 529–532. (doi:10.1007/s00439-005-1281-4)
44 Gusmão, L. *et al.* 2005 Mutation rates at Y chromosome specific microsatellites. *Hum. Mutat.* **26**, 520–528. (doi:10.1002/humu.20254)
45 Kayser, M. *et al.* 2000 Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* **66**, 1580–1588. (doi:10.1086/302905)
46 Bocquet-Appel, J.-P., Naji, S., Linden, M. V. & Kozlowski, J. K. 2009 Detection of diffusion and contact zones of early farming in Europe from the space-time distribution of 14C dates. *J. Archaeol. Sci.* **36**, 807–820. (doi:10.1016/j.jas.2008.11.004)
47 Sjödin, P. & François, O. 2011 Wave-of-advance models of the diffusion of the Y chromosome haplogroup R1b1b2 in Europe. *PLoS ONE* **6**, e21592. (doi:10.1371/journal.pone.0021592)

European Journal of Human Genetics (2009) 17, 693 – 696

www.nature.com/ejhg

SHORT REPORT

The Etruscan timeline: a recent Anatolian connection

Francesca Brisighelli[1,2], Cristian Capelli[3], Vanesa Álvarez-Iglesias[1], Valerio Onofri[4], Giorgio Paoli[5], Sergio Tofanelli[5], Ángel Carracedo[1,6], Vincenzo L Pascali[2] and Antonio Salas*[,1]

[1]*Unidade de Xenética, Facultade de Medicina, Instituto de Medicina Legal, Universidade de Santiago de Compostela, Galicia, Spain;* [2]*Forensic Genetics Laboratory, Institute of Legal Medicine, Università Cattolica del Sacro Cuore, Rome, Italy;* [3]*Department of Zoology, University of Oxford, Oxford, UK;* [4]*Institute of Legal Medicine, Universita' Politecnica delle Marche, Policlinico Torrette, Ancona, Italy;* [5]*Department of Biology, University of Pisa, Pisa, Italy;* [6]*Fundación Pública Galega de Medicina Xenómica (FPGMX-SERGAS), CIBER enfermedades raras, Santiago de Compostela, Galicia, Spain*

The origin of the Etruscans (the present day Tuscany, Italy), one of the most enigmatic non-Indo-European civilizations, is under intense controversy. We found novel genetic evidences on the mitochondrial DNA (mtDNA) establishing a genetic link between Anatolia and the ancient Etruria. By way of complete mtDNA genome sequencing of a novel autochthonous Tuscan branch of haplogroup U7 (namely U7a2a), we have estimated an historical time frame for the arrival of Anatolian lineages to Tuscany ranging from 1.1 ± 0.1 to 2.3 ± 0.4 kya B.P.

European Journal of Human Genetics (2009) **17,** 693–696; doi:10.1038/ejhg.2008.224; published online 3 December 2008

Keywords: mtDNA; haplogroup; complete genome; Tuscany; Etruscans

Introduction

The origin of the Etruscans, one of the most ancient and enigmatic non-Indo-European civilizations, is being the target of a controversial debate. A recent study identified among modern Tuscans a rather high prevalence of Near Eastern mtDNA haplogroups and an exclusive haplotype sharing between them and Near Eastern populations.[1] The finding has been interpreted as evidence in support of the classical theory that Etruscans may have come from the East through the Mediterranean Sea (Herodotus, *Historiae*, Vol I, p 94), which currently find little support by archaeologists and historians.[2] In favor of the Eastern Mediterranean origin of the Etruscan civilization, the finding that the extent of mtDNA variation observed in Tuscan cattle breeds is similar to that observed in the Near East and much higher than that observed in the rest of Italy and Europe.[3] The two facts could be compliant with other hypotheses. Thus, studies on fossil DNA in Italy have identified ancient pre-Neolithic bovine – aurochs – whose types are closer to modern bovine than West European aurochs: this contradicts the bovine migration theory and suggests either *in loco* domestication or population continuity across Italy–Balkans–Anatolia during the Palaeolithic[4] (however, see the recent study by Achilli *et al*[5] for a high-resolution study on mitochondrial DNA (mtDNA) of aurochs and domestic cattle). Furthermore, currently available analysis of archaeological Etruscan remains seems to indicate genetic continuity with Tuscans, with closer, but not specific, affinity with Anatolia[6] (however, see Bandelt[7] and Malyarchuk and Rogozin[8]).

To further test these hypotheses, we have analyzed a total of 258 Tuscan samples using mtDNA single nucleotide polymorphisms (SNPs), which allow the classification of Near Eastern typical haplogroups (HV lineages that are non-H and non-HV0, R0a, U7 and U3). mtDNA complete genome sequencing of a novel and autochthonous U7 sub-

***Correspondence: Dr A Salas, Unidade de Xenética, Facultad de Medicina, Instituto de Medicina Legal, Universidad de Santiago de Compostela, 15782, Galicia, Spain.**
Tel.: +34 981 582 327; Fax: +34 981 580 336;
E-mail: antonio.salas@usc.es
Received 14 April 2008; revised 14 October 2008; accepted 15 October 2008; published online 3 December 2008

haplogroup has allowed, for the first time, to provide a time frame for this event.

Materials and Methods

Samples

We have undertaken a sample collection campaign covering 10 areas in Tuscany,[9] whose geographical location extends to a wide area covering continental Etruria and the Elba Island (whose ferrous beds exploitation has set a landmark in driving ancient Etruscan craftsmanship): Arezzo ($N=11$), Chiusi ($N=36$), Collevecchio ($N=24$), Elba Island ($N=53$), Magliano Sabina ($N=49$), Monte Fiascone ($N=17$), Pitigliano ($N=16$), Tarquinia ($N=15$), Tuscania ($N=26$) and Vulci ($N=11$).

Genotyping of mtDNA SNPs

We have used the minisequencing technique for screening a total of 258 samples,[10] all for a set of 24 mtDNA SNPs that allow to classify mtDNA sequences into major European haplogroups plus those that are more likely to be of Near East origin (Supplementary Table S1). Those mtDNAs with a SNP profile compatible with a typical Near Eastern haplogroups, that is, HV lineages that are non-H and non-HV0, R0a, U7 and U3, were further sequenced for the first hypervariable segment (HVS-I); only a small fraction of them (~ 10 of the total sample size) were finally confirmed to belong to the mentioned haplogroups. Some other samples showing ambiguous haplogroup affiliation (eg, members of the broad macro-haplogroup N*) or that could reveal some phylogeographic information at the control region level (eg, haplogroups I, W, X) were also sequenced for the HVS-I. In total, 63 samples out of 258 were sequenced for the HVS-I.

Automatic sequencing

PCR amplification was carried out in a 9700 Thermocycler (AB). The temperature profile for 32 cycles of amplification was 95°C for 10 s, 60°C for 30 s and 72°C for 30 s. Sequencing primers were described earlier by Wilson *et al*.[11] PCR product purification and sequencing were performed according to Salas *et al*.[12]

Nine samples from the Isle of Elba were sequenced for the complete mtDNA genome. The primers used for PCR amplification and sequencing were those reported by Torroni *et al*[13] with minor modifications. More technical details concerning the PCR and sequencing reaction can be provided under request.

A posteriori sequence quality was evaluated following the methods described earlier.[14–17]

Databases

We have compiled a total of 15 631 HVS-I sequences into a database that contains 13 155 West European profiles (including 1099 from different areas of Italy) and 2476 from Near East.

Coalescence age

Estimation of the time to the most recent common ancestor of each cluster and SDs was carried out according to Saillard *et al*[18] and using an evolutionary rate estimate of $1.26 \times 0.08 \times 10^{-8}$ base substitutions (other than a deletion or insertion) per nucleotide per year in the coding region (between 577 and 16 023), corresponding to 5140 years per substitution in the entire coding region.[19] The coalescence age needed to accumulate the variation within U7a2a was estimated using both control and coding region information, which corresponds in Figure 1 with the first and the second term, respectively.

Results and discussion

A total of 63 mtDNAs were sequenced for the HVS-I as described in Materials and Methods. The resulting haplotypes were searched across a European and Near Eastern database of more than 15 500 sequences. Overall, 34 out of the 63 sequenced Tuscan individuals (21 haplotypes) have a counterpart in Near East. Five HVS-I haplotypes (eight individuals who constitute $\sim 3\%$ of the individuals in the total sample) singled out of the 63 sequenced individuals were not present in a large European database containing over 12 500 and including more than 1000 Italian sequences from outside Tuscany. Interestingly, some of those 'Near Eastern sequences' emerging from our Tuscan sample did match with the Tuscan haplotypes described by Achilli *et al*.[1]

On the basis of combined information of SNPs and HVS-I sequence data, we confirmed that $\sim 10\%$ (26 individuals out of 258) of our Tuscans actually belong to one of the typical Near Eastern haplogroups (see Supplementary Table S1), and have also a match with Near East populations. All of these Near East haplotypes are diverse (with the exception of those belonging to U7, see below) and fall at the tips of the phylogeny, suggesting a recent arrival to the region.

The typical Near Eastern U7 haplogroup occurs at relatively high frequency in the Elba Island ($\sim 17\%$; 9 mtDNAs out of 53), and all of these U7 mtDNAs share the same HVS-I motif (T16271C-A16318T-T16519C), indicating that this lineage could represent a Near Eastern founder in the Isle. The T16271C-A16318T motif matches only two additional sequences in a worldwide database of more than $>70\,000$ profiles; interestingly, both correspond to DNA samples collected in the 'Etruscan area' (in Lucca; author's personal communication; GenBank acc. no.: DQ081609 and DQ081665;[20]). Complete genome sequencing of these nine U7 mtDNAs allowed the identification of a new subclade, U7a2a, characterized by transitions A13395G and T16271C. U7a2a is a sub-branch of U7a2; to our knowledge, only two other U7a2 complete genomes lacking the diagnostic motif of U7a2a have been reported in the literature, one was identified in a Pakistani and the other in an Andalusian (see Figure 1). The amount of variation

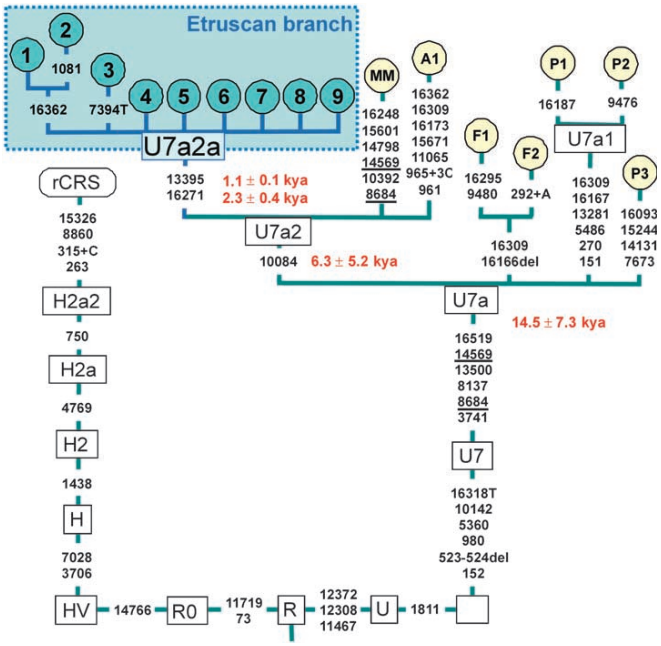


Figure 1 Phylogeny of complete genomes belonging to haplogroup U7. Source of the data: nos. 1–9, present study (GenBank acc. no. EU445683-EU445691; MM = no. AndalAF_11 from Andalucia, Southern Spain (acc. no. AF382011); F1, no. 146 from Finland (acc. no. AY339547); F2, no. 147 from Finland (acc. no. AY339548); P1, no. B19 from India (acc. no. AY714013); P2, no. B81 from India (acc. no. AY714014); P3, no. C22 from India (acc. no. AY714004); A1, no. 13 from Pakistan (A Achilli, personal communication; acc. no. AY882391). The coalescence age needed to accumulate the variation within U7a2a was estimated using both control (top) and coding (bottom) region information.

accumulated within U7a2a Etruscan cluster (assuming a single founder) can be dated in the range 1.1 ± 0.1 to 2.3 ± 0.4 kya B.P., consistent with a recent arrival of this haplogroup to this Isle and compatible with the Etrurian culture (9th–1st century BC).

The investigation of a large and representative sample set and the analysis of complete mtDNA genomes support the hypothesis that Tuscany still preserves the fingerprint of a historical connection with the Near East. However, it should be stressed that this represents just a minor component of the Tuscan genetic make-up and suggests that historically different layers were superimposed over the Mesolithic gene pool of the Peninsula.

Note added in proof

Analysis performed by coalescent simulations[21] suggested a model with little or no continuity between Ancient Etruscans and Modern Tuscans. However, the ancient dataset was extremely small and only a larger sample size would set the issue of diachronic continuity in Tuscany.

Acknowledgements

We would like to thanks Antonio Torroni for his useful comments and suggestions on the article and Alessandro Achilli for his help in submitting the complete genome fasta files to GenBank. The grant of the Xunta de Galicia (PGIDIT06PXIB208079PR), and the grant from the Fundación de Investigación Médica Mutua Madrileña, given to AS, partially supported this project.

References

1 Achilli A, Olivieri A, Pala M *et al*: Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. *Am J Hum Genet* 2007; **80**: 759–768.
2 Pallottino M: *Storia della Prima Italia*. Rusconi: Italy, Milano, 1999.
3 Pellecchia M, Negrini R, Colli L *et al*: The mystery of Etruscan origins: novel clues from *Bos taurus* mitochondrial DNA. *Proc Biol Sci* 2007; **274**: 1175–1179.

4 Beja-Pereira A, Caramelli D, Lalueza-Fox C *et al*: The origin of European cattle: evidence from modern and ancient DNA. *Proc Natl Acad Sci USA* 2006; **103**: 8113–8118.
5 Achilli A, Olivieri A, Pellecchia M *et al*: Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr Biol* 2008; **18**: R157–R158.
6 Vernesi C, Caramelli D, Dupanloup I *et al*: The Etruscans: a population-genetic study. *Am J Hum Genet* 2004; **74**: 694–704.
7 Bandelt H-J: Etruscan artifacts. *Am J Hum Genet* 2004; **75**: 919–920; author reply 923–917.
8 Malyarchuk BA, Rogozin IB: On the Etruscan mitochondrial DNA contribution to modern humans. *Am J Hum Genet* 2004; **75**: 920–923; author reply 923–927.
9 Capelli C, Brisighelli F, Scarnicci F *et al*: Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. *Mol Phylogenet Evol* 2007; **44**: 228–239.
10 Alvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A: Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. *Forensic Sci Int Genet* 2007; **1**: 44–55.
11 Wilson MR, Polanskey D, Butler J, DiZinno JA, Replogle J, Budowle B: Extraction, PCR amplification and sequencing of mitochondrial DNA from human hair shafts. *Biotechniques* 1995; **18**: 662–669.
12 Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo Á: mtDNA analysis of the Galician population: a genetic edge of European variation. *Eur J Hum Genet* 1998; **6**: 365–375.
13 Torroni A, Rengo C, Guida V *et al*: Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* 2001; **69**: 1348–1356.
14 Bandelt H-J, Quintana-Murci L, Salas A, Macaulay V: The fingerprint of phantom mutations in mitochondrial DNA data. *Am J Hum Genet* 2002; **71**: 1150–1160.
15 Bandelt H-J, Salas A, Bravi CM: Problems in FBI mtDNA database. *Science* 2004; **305**: 1402–1404.
16 Salas A, Bandelt H-J, Macaulay V, Richards MB: Phylogeographic investigations: the role of trees in forensic genetics. *Forensic Sci Int* 2007; **168**: 1–13.
17 Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt H-J: A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. *Biochem Biophys Res Commun* 2005; **335**: 891–899.
18 Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S: mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* 2000; **67**: 718–726.
19 Mishmar D, Ruiz-Pesini E, Golik P *et al*: Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci USA* 2003; **100**: 171–176.
20 Falchi A, Giovannoni L, Calo CM *et al*: Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms. *J Hum Genet* 2006; **51**: 9–14.
21 Belle EM, Ramakrishnan U, Mountain JL, Barbujani G: Serial coalescent simulations suggest a weak genealogical relationship between Etruscans and modern Tuscans. *Proc Natl Acad Sci USA* 2006; **103**: 8012–8017.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

European Journal of Human Genetics (2009) 17, 848 – 852

www.nature.com/ejhg

SHORT REPORT

Moors and Saracens in Europe: estimating the medieval North African male legacy in southern Europe

Cristian Capelli*[,1], Valerio Onofri[2], Francesca Brisighelli[3,4], Ilaria Boschi[4], Francesca Scarnicci[4], Mara Masullo[4], Gianmarco Ferri[5], Sergio Tofanelli[6], Adriano Tagliabracci[2], Leonor Gusmao[7], Antonio Amorim[7,8], Francesco Gatto[9], Mirna Kirin[10], Davide Merlitti[11], Maria Brion[3], Alejandro Blanco Verea[3], Valentino Romano[12], Francesco Cali[13] and Vincenzo Pascali[4]

[1]*Department of Zoology, University of Oxford, Oxford, UK;* [2]*Institute of Legal Medicine, Universita' Politecnica delle Marche, Policlinico Torrette, Ancona, Italy;* [3]*Medicine Genomic Group, Hospital-University complex of Santiago (CHUS), University of Santiago de Compostela, Spain;* [4]*Instituto di Medicina Legale, Universita' Cattolica del S. Cuore, Rome, Italy;* [5]*Department of Diagnostic and Laboratory Service and Legal Medicine, Section of Legal Medicine, University of Modena and Reggio Emilia, Italy;* [6]*Department of Biology, Anthropology Unit, University of Pisa, Italy;* [7]*IPATIMUP – Institute of Molecular Pathology and Immunology of the University of Porto, Portugal;* [8]*Faculty of Sciences, University of Porto, Portugal;* [9]*Biotechnology Unit, Istituto Zooprofilattico Sperimentale Lazio e Toscana, Rome, Italy;* [10]*Public Health Sciences, University of Edinburgh, Edinburgh, Scotland;* [11]*Scuola Normale Superiore di Pisa, Pisa, Italy;* [12]*Dipartimento di Oncologia Sperimentale e Applicazioni Cliniche Università di Palermo, Italy;* [13]*Oasi Institute for Research on mental Retardation and Brain Aging (IRCCS), Troina, Italy*

To investigate the male genetic legacy of the Arab rule in southern Europe during medieval times, we focused on specific Northwest African haplogroups and identified evolutionary close STR-defined haplotypes in Iberia, Sicily and the Italian peninsula. Our results point to a higher recent Northwest African contribution in Iberia and Sicily in agreement with historical data. southern Italian regions known to have experienced long-term Arab presence also show an enrichment of Northwest African types. The forensic and genomic implications of these findings are discussed.
European Journal of Human Genetics (2009) **17,** 848–852; doi:10.1038/ejhg.2008.258; published online 21 January 2009

Keywords: Y chromosome; North Africa medieval legacy; southern Europe

Introduction

After the collapse of the Roman Empire in Europe, the Arab dominance across the Mediterranean was one of the most impressive historical events that occurred in this region. Arabs appeared on the southern shores of the Mediterranean in the early seventh century and quickly conquered North Africa. They spread their language and religion to the native Northwest (NW) African Berber populations, which represented the bulk of the Muslim army that later conquered southern Europe.[1,2] Referred to either as Moors (in Iberia) or Saracens (in South Italy and Sicily), their arrival in Europe dates to 711 AD, rapidly subduing most of Iberia and Sicily (831 AD). Among European kingdoms their presence was seen as a constant danger, and only by the fifteenth century was the Iberian reconquest completed.[3] In the thirteenth century Frederick II destroyed

***Correspondence: Dr C Capelli, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.**
Tel: +44 1865 271261; Fax: +44 1865 310447;
E-mail: cristian.capelli@zoo.ox.ac.uk
Received 27 May 2008; revised 27 November 2008; accepted 2 December 2008; published online 21 January 2009

Arab rule in Sicily and between 1221 and 1226 he moved all the Arabs of Sicily to the city of Lucera, north of Apulia.[3] Lucera was later destroyed by Charles II (1301) but an Arab community was recorded in Apulia in 1336.[3] Guerrilla warfare was still conducted by Arabs in Sicily even after Frederick II's actions.[3]

So far, Y chromosome studies attempting to estimate the medieval North African (MNA) contribution to southern Europe have focused almost exclusively on the North African haplogroup E3b1b1b-M81, and have only partially taken into consideration the evolutionary relationships among haplotypes.[4–7] To generate a more comprehensive view of the genetic legacy of the MNA dominance in Europe, we systematically screened for Y chromosome haplotypes within three NW African specific haplogroups, across multiple southern European populations, and performed additional genotyping to refine the available genetic data. Our results confirm a general correlation between historical and genetic data: Iberia and Sicily are the regions with the highest MNA male legacy.

Materials and methods

Identification of recently introgressed NW African haplotypes

Given the historical indication of a prevalently Berber origin for the Arab groups invading southern Europe,[2,3] we focused on NW African specific haplogroups as markers of MNA contribution to this region. Haplogroups E1b1b1b (M81 derived), E1b1b1a-β (M78 derived chromosomes showing the rare DYS439 allele 10) and a subset of J1 (M267 derived) were identified in the literature as being NW Africa specific, together accounting for between 58 and 90% of males in populations from this area, but never above 13% in Europe.[8–11] We note that the other lineages present in these populations would also have been brought over to Europe, and any account of the total MNA contribution to present day Europe should take these into consideration.

Given a number of investigated loci n, and a mutation rate μ (estimated using locus specific data as in reference[12]), it is possible to obtain the posterior distribution of the Time to the Most Recent Common Ancestor for any pair of haplotypes differing at k loci, using the approach implemented in reference.[13] The selected method is based on the infinite alleles model, a reasonable approximation when few mutations are expected to occur, as in the temporal framework evaluated here. So, considering 9 loci and 40 generations (approximately 1200 years ago with a 31-year generation length[14]), either 0 or 1 mutational difference is the most likely consequence. Two mutations are only slightly less likely, but overlap with other much more ancient events, for example 80 generations or 2400 years ago. Posterior distributions for more ancient events have probability peaks centred on a higher number of differences, with 0–1 mutations being extremely unlikely (data not shown). Therefore, following this, European Y chromosomes within the three haplogroups identical to, or with one mutational difference from, NW African STR haplotypes were considered compatible with an MNA ancestry. In Iberia and peninsular Italy, they account for 90, 78 and 42% of the E1b1b1b, E1b1b1a-β and J1 chromosomes respectively.

Samples

A NW African database was constructed for haplotype comparisons including more than 400 samples genotyped at nine STR loci (DYS19, DYS389 I–II, DYS390, DYS391, DYS392, DSY393, and the bi-allelic DYS385). The database included 127 Berbers from Tunisia;[15,16] 102 South Tunisians;[17] 109 Moroccan Arab and Berber speakers;[18] 50 Moroccan and 52 Tunisians (unpublished data). NW African specific haplogroups were identified by further genotyping of samples that were previously described elsewhere.[5–7,19–21] We also included a Basque dataset[22,23] and two novel Italian samples (Lucera and Veneto; Table 1). Within these populations, all E1b1b1a chromosomes were scored for the DYS439 locus to identify the E1b1b1a-β cluster[9] and the M267 marker was investigated in those chromosomes previously identified as J*(xJ2). Alternatively, the DYS458 .2 allele was used to identify the J1 types within J*(xJ2) chromosomes.[24] All the individuals within E1b1b1b, E1b1b1a-β and J1 were also genotyped for the same nine STRs as the NW Africans (DYS19, DYS389 I–II, DYS390, DYS391, DYS392, DYS393 and DYS385). The DYS385 bilocal locus was considered as two different loci, the smaller allele assigned to locus DYS385a and the larger to DYS385b. A previous investigation[25] showed that misassignment would influence only a minimal fraction of the haplotypes and so this can be assumed to have a negligible effect on our estimates. A Sicilian population was also included (samples overlapping in references[26,27]). Sicilian genotypes were screened for E1b1b1* and J*(xJ2) lineages, and did not include DYS439. Within the E1b1b1* and J*(xJ2) haplogroups, 8 and 3 chromosomes, respectively, were found close to NW African types. These samples were then made available for further genotyping, to include DYS439, M78, M81 and M267. We note that because of partial sampling across NW Africa, a subset of the European chromosomes with true MNA ancestry could potentially fail to be identified. However, given the general homogeneity observed across NW Africa, the number of populations included, and the large dataset used, we believe that this is unlikely to influence our results.

Results and discussion

To address the degree of historical NW African contribution, we used a combined SNP-STR approach. The coalescent times for the three NW African specific haplogroups

Table 1 Historically introduced NW African types in Italy and Iberia

	Sample	n	*E1b1b1b*	*E1b1b1a-β*[a]	*J1*	*Total %*
1	Val Badia	34	0.0	0.0	0.0	0.0
2	Veneto	55	1.8	0.0	0.0	1.8
3	Central Emilia	62	0.0	0.0	0.0	0.0
4	Central-Tuscany	41	0.0	0.0	2.4	2.4
5	Tuscany-Latium border	79	0.0	0.0	0.0	0.0
6	North-East Latium	55	1.8	0.0	0.0	1.8
7	Marche	221	0.0	0.5	0.9	1.4
8	South Latium	51	0.0	0.0	0.0	0.0
9	East Campania	84	2.4	1.2	1.2	4.8
10	North-West Apulia	46	4.3	0.0	2.2	6.5
11	Lucera	60	1.7	1.7	0.0	3.3
12	West Calabria	56	0.0	0.0	0.0	0.0
13	South Apulia	71	0.0	0.0	1.4	1.4
	Peninsular Italy	915	0.8	0.3	0.7	1.7
14	Sicily	93	2.2	2.2	3.2	7.5
15	Portugal[b]	659	5.0	0.3	1.8	7.1
16	Galicia[c]	292	4.1	0.7	2.1	6.8
17	Cantabria[c]	161	13.0	3.1	2.5	18.6
18	Basques[d]	168	0.6	0.0	0.6	1.2
19	Basques[e]	43	2.3	0.0	0.0	2.3
20	Catalans[e]	16	0.0	0.0	0.0	0.0
21	Andalusians[e]	37	5.4	0.0	0.0	5.4
	Total Spain	717	5.2	1.0	1.5	7.7
	Total Iberia	1376	5.1	0.7	1.7	7.4

Frequencies of E1b1b1b, E1b1b1a-β and J1 chromosomes with 0-1-steps neighbour chromosome within the NW African dataset; the first column refers to the geographic location in Figure 1.
[a]E1b1b1a-β chromosomes were identified as M78 derived bearing the DYS439 allele 10[9].
[b]Overlapping with Beleza *et al*.[7]
[c]Samples from Brion *et al*;[6] a subset of the J and E samples have been further tested with M81, M78 and DYS439 and used to estimate frequencies. J1 samples have been identified as J samples with the 0.2 DYS458 allelic variant.[22]
[d]Combined data from Alonso *et al*;[22] Garcia *et al*.[21]
[e]DYS439 and DYS385 were genotyped in the relevant samples from Bosch *et al*,[5] except for one Basque sample, not included.

ranges between 5000 and 24 000 years, spanning a number of historical scenarios each potentially explaining their presence on the Northern Mediterranean shores.[9,10] It follows that estimating MNA genetic legacy on the basis of haplogroups' occurrence only would be misleading. To avoid this limitation, we have extended our analysis to include STR data whose high mutation rate allows one to focus on more recent events. We screened more than 2300 South European samples (Figure 1; Table 1) to identify those haplotypes which are evolutionary close to NW African chromosomes. Total frequencies for these chromosomes range between 0 and 19% across southern Europe, the highest being in Cantabria and comprising a sample from the Pas Valley, previously shown to have an extremely high frequency of the North African haplogroup E1b1b1b.[9] Our estimates of NW African chromosome frequencies were highest in Iberia and Sicily, in accordance with the long-term Arab rule in these two areas.[3] The chromosome frequencies in the two samples were not significantly different from each other (Fisher's exact test $P=0.83$) but were both significantly different from the peninsular Italy sample ($P<0.01$). An inspection of Table 1 reveals a non-random distribution of MNA types in the Italian peninsula, with at least a twofold increase over the Italian average estimate in three geographically close samples across the southern Apennine mountains (East Campania, Northwest Apulia, Lucera). When pooled together, these three Italian samples displayed a local frequency of 4.7%, significantly different from the North and the rest of South Italy ($P<0.01$), but not from Iberia and Sicily ($P=0.12$ and $P=0.33$, respectively). Arab presence is historically recorded in these areas following Frederick II's relocation of Sicilian Arabs.[3] In Iberia, a non-random distribution might also potentially be present, as suggested by our lower estimates in the northeast (Basque region and Catalans), but more samples across the peninsula will be required to properly address this issue. Assuming that a large population in regions such as Iberia, Sicily and Italy was present in the past, the ratio between Y chromosomes with a MNA ancestry and other types will have stayed approximately constant across time. Smaller areas, however, would have been influenced by drift, in the Pas Valley for example. Consistent with historical data,[3] no population in Central Europe or the Balkans shows the presence of recently introgressed NW African types[9,10,28] besides a few chromosomes in Albania and Romania.[29]

The increasing use of highly structured distributions of Y chromosome types to investigate the ethnic/geographic

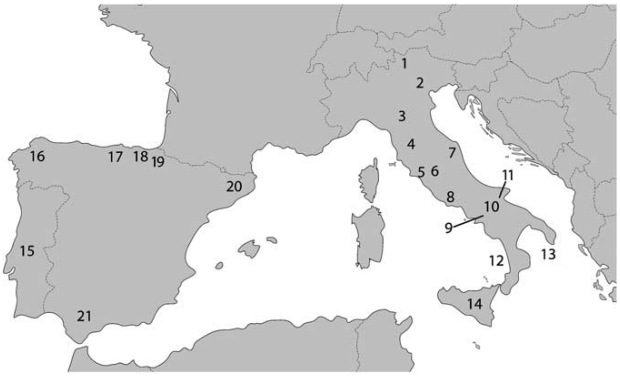


Figure 1 Geographical location of the investigated southern European samples. Numbers are same as in Table 1.

origin of unknown samples[30] gives the identification of regions in Italy enriched with recently introgressed NW African types forensic relevance. We found that more than 56% of the Italian individuals identified here as having a recent NW African do not have a match in a large Italian Y chromosome dataset comprising almost 1200 individuals.[31] Of these, 31% instead perfectly overlap with types from NW African populations, potentially providing misleading advice to investigators. Such results are also of interest in the light of the expanding business of genealogical services offering Y chromosome analysis to identify an individual's ethnic ancestry. Our results clearly confirm that conclusions based on single chromosomes should be taken very cautiously.[32] What are the expected genomic consequences of this historically recent admixture event? Suppose that 40 generations ago there was a 5% male introgression of African DNA into the European gene pool, corresponding to a total contribution of 2.5% of genetic material. Immediately after the admixture event, a fraction of chromosomes within Europe would have African ancestry. Recombination since this event will have substantially reduced the size of the fragments of African ancestry within European haplotypes, and with these parameters we would today expect to see an approximately exponential distribution (measuring size using genetic distance) of fragment sizes, with a mean value of roughly 2.6 cM. Assuming a genome-wide average recombination rate of 1.3 cM/Mb,[33] 2.5% of a typical present day southern European genome would consist on average of 2 Mb regions of African DNA. We therefore believe that signatures of this event would be correctly identified using modern dense genotype data.[34] By using northern Italian and Mozabite samples recently genotyped for a large SNP autosomal dataset[35] as the best available proxy of Italian and northern African populations, we estimated that about 41.5% of more than 640 000 genotyped SNPs showed an absolute allele frequency difference of at least 10% between the two groups. Such frequency differences (and sometimes even smaller) between cases and controls characterized the vast majority of the inferred disease-causing SNPs in a recent genome-wide investigation.[36] In general then, it is critical to take population structure into account so as to avoid false positives in case–control association studies.[37] Thus, an understanding of similar historical admixture events is likely to aid researchers conducting such studies.

Acknowledgements

We thank Elena Bosch and Walther Parson for kindly providing unpublished data; Giovanni Destro-Bisol for commenting a preliminary version of the article; Dr Trincucci for support in the sampling of the Lucera inhabitants; Marcello Menegatti, Cristian Sossai and the Associazione Culturale 'Borghi dell'Ovest' for the Veneto samples. CC thanks Simon Myers and Garrett Hellenthal for comments and suggestions on the genomic structure implication following recent admixture events, Jim Wilson for support and Prof Francesco Sabatini for discussion on the history of Lucera. CC is a RCUK Academic Fellow.

References

1 Davies RHC: *A History of Medieval Europe*. London, UK: Longmann Group Limited, 1988, pp 83–101.
2 Hitti P: *The Arabs: A Short History*. Washington DC: Gateway, 1990.
3 Norman D: *The Arabs and Medieval Europe*. London, UK: Longmann Group Limited, 1975.
4 Rosser ZH, Zerjal T, Hurles ME *et al*: Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* 2000; **67**: 1526–1543.
5 Bosch E, Calafell F, Comas D *et al*: 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.
6 Brion M, Quintans B, Zarrabeitia M *et al*: Micro-geographical differentiation in Northern Iberia revealed by Y-chromosomal DNA analysis. *Gene* 2004; **329**: 17–25.

7 Beleza S, Gusmão L, Lopes A *et al*: Micro-phylogeographic and demographic history of Portuguese male lineages. *Ann Hum Genet* 2006; **70**: 181–194.
8 Arredi B, Poloni ES, Paracchini S *et al*: A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet* 2004; **75**: 338–345.
9 Cruciani F, La Fratta R, Santolamazza P *et al*: Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Genet* 2004; **74**: 1014–1022.
10 Semino O, Magri C, Benuzzi G *et al*: Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet* 2004; **74**: 1023–1034.
11 History and geography of human Y-chromosome in Europe: a SNP perspective Paolo Francalacci & Daria Sanna. *Journal of Anthropological Sciences (J Anthropol Sci)* 2008; **86**: 59–89.
12 Walsh B: Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics* 2001; **158**: 897–912.
13 Gusmão L, Sánchez-Diz P, Calafell F *et al*: Mutation rates at Y chromosome specific microsatellites. *Hum Mutat* 2005; **26**: 520–528.
14 Helgason A, Hrafnkelsson B, Gulcher JR, Ward R, Stefansson K: A population wide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am J Hum Genet* 2003; **72**: 1370–1389.
15 Frigi S, Pereira F, Pereira L *et al*: Data for Y-chromosome haplotypes defined by 17 STRs (AmpFLSTR® Yfiler™) in two Tunisian Berber communities. *Forensic Sci Int* 2006; **160**: 80–83.
16 Cherni L, Pereira L, Goios A *et al*: Y-chromosomal STR haplotypes in three ethnic groups and one cosmopolitan population from Tunisia. *Forensic Sci Int* 2005; **152**: 95–99.
17 Ayadi I, Ammar-Keskes L, Rebai A: Haplotypes for 13 Y-chromosomal STR loci in South Tunisian population (Sfax region). *Forensic Sci Int* 2006; **164**: 249–253.
18 Quintana-Murci L, Bigham A, Rouba H *et al*: Y-chromosomal STR haplotypes in Berber and Arabic-speaking populations from Morocco. *Forensic Sci Int* 2004; **140**: 113–115.
19 Onofri V, Alessandrini F, Turchi C *et al*: Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis. *Int J Legal Med* 2007; **121**: 234–237.
20 Capelli C, Brisighelli F, Scarnicci F *et al*: Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. *Mol Phylogenet Evol* 2007; **44**: 228–239.
21 Ferri G, Alù M, Corradini B *et al*: Slow and fast evolving markers typing in Modena males (North Italy). *Forensic Sci Int Genet*, (in press).
22 García O, Martín P, Gusmão L *et al*: A Basque Country autochthonous population study of 11 Y-chromosome STR loci. *Forensic Sci Int* 2004; **145**: 65–68.
23 Alonso S, Flores C, Cabrera V *et al*: The place of the Basques in the European Y-chromosome diversity landscape. *Eur J Hum Genet* 2005; **13**: 1293–1302.
24 Myres NM, Ekins JE, Lin AA *et al*: 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.
25 Niederstätter H, Berger B, Oberacher H, Brandstätter A, Huber CG, Parson W: Separate analysis of DYS385a and b versus conventional DYS385 typing: is there forensic relevance? *Int J Legal Med* 2005; **119**: 1–9.
26 Capelli C, Redhead N, Romano V *et al*: Population structure in the Mediterranean basin: a Y chromosome perspective. *Ann Hum Genet* 2006; **70**: 207–225.
27 Robino C, Inturri S, Gino S *et al*: Y-chromosomal STR haplotypes in Sicily. *Forensic Sci Int* 2006; **159**: 235–240.
28 Berger B, Lindinger A, Niederstätter H *et al*: Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay. *Int J Legal Med* 2005; **119**: 241–246. Erratum in: Int J Legal Med 2006; 120: 255.
29 Bosch E, Calafell F, González-Neira A *et al*: Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Ann Hum Genet* 2006; **70**: 459–487.
30 Wetton JH, Tsang KW, Khan H: Inferring the population of origin of DNA evidence within the UK by allele-specific hybridization of Y-SNPs. *Forensic Sci Int* 2005; **152**: 45–53.
31 Presciuttini S, Caglià A, Alù M *et al*: Y-chromosome haplotypes in Italy: the GEFI collaborative database. *Forensic Sci Int* 2001; **122**: 184–188.
32 King TE, Parkin EJ, Swinfield G *et al*: Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. *Eur J Hum Genet* 2007; **15**: 288–293.
33 Yu A, Zhao C, Fan Y *et al*: Comparison of human genetic and sequence-based physical maps. *Nature* 2001; **409**: 951–953.
34 Frazer KA, Ballinger DG, Cox DR *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
35 Li JZ, Absher DM, Tang H *et al*: Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008; **319**: 1100–1104.
36 Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.
37 Marchini J, Cardon LR, Phillips MS *et al*: The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36**: 512–517.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

Available online at www.sciencedirect.com
SCIENCE DIRECT
Forensic Science International 159 (2006) 64–70
Forensic Science International
www.elsevier.com/locate/forsciint

Announcement of population data

A 9-loci Y chromosome haplotype in three Italian populations☆

Cristian Capelli [a,b,*], Barbara Arredi [a], Laura Baldassari [a], Ilaria Boschi [a], Francesca Brisighelli [a], Alessandra Caglià [a,1], Marina Dobosz [a], Francesca Scarnicci [a], Giuseppe Vetrugno [a], Vincenzo L. Pascali [a]

[a] *Laboratorio di Genetica Forense, Istituto di Medicina Legale, Università Cattolica del S. Cuore, Lgo F. Vito 1, Rome, Italy*
[b] *Promega Italia, Via Decembrio 28, 20137 Milano, Italy*

Received 25 January 2005; received in revised form 3 May 2005; accepted 10 May 2005
Available online 5 July 2005

Abstract

Three geographic areas of Italy have been sampled and genotyped for 9 Y chromosome STRs: DYS19, DYS385, DYS388, DYS389 I and II, DYS390, DYS391, DYS392, DYS393. Sampling was focused on residents of small areas, well distant from major urban centres. Only individuals whose grandfather would live in the same area were included. A total of 210 unrelated individuals were collected. Distribution of genetic variation across the three samples and comparison with previously published Italian database indicated that so far Y chromosome diversity has been only partially explored in the Italian Peninsula.

Keywords: Y chromosome; STRs; Italian populations

1. Population

Three areas of the Italian peninsula were selected for genomic sampling: South Apulia (SA), Marche (MA) and a broader area comprising South of Tuscany and the North of Latium (TL) (Fig. 1). The sampling sizes were 72 (SA), 59 (MA) and 79 (TL). Individuals were paternally unrelated, they had different surnames and they would define themselves as belonging to a paternal lineage residing in the area from at least three generations.

2. Extraction

Samples were collected by buccal swabs. Extraction was performed with a modified salting-out method [1] readapted to buccal cells. Swabs were incubated in 500 μl of 0.2 M sodium acetate, 35 μl of 10% SDS and 20 μl of 20 mg/ml Proteinase K for 16 h at 56 °C. They were then removed and 500 μl of a 3 M NaCl solution was added. Proteins were removed by centrifugation, and the DNA precipitated by adding 1 ml of ethanol 100% at −20 °C for few hours. After centrifugation, DNA pellet was washed with ethanol 70% twice, dried and re-suspended in water.

3. Genotyping

Samples were genotyped by two in-house developed multiplexes: one contained DYS389 I and II, DYS390 and

☆ The opinions expressed in this article by C. Capeli do not necessarily reflect those of Promega Corporation.

* Corresponding author. Present address: Genetic Identity Europe, Promega Corporation, Madison, WI, USA. Tel.: +39 0773 458535; fax: +39 0773 458535.
E-mail address: cristian.capelli@promega.com (C. Capelli).
[1] Present address: DCPC, Servizio di Polizia Scientifica Area Biologica, Viale dell'Aereonautica, 7-00144 Rome, Italy.

0379-0738/$ – see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2005.05.026

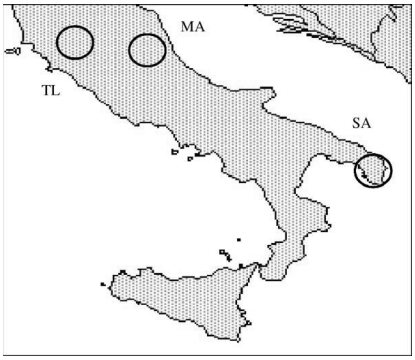


Fig. 1. Geographic localisation of the three sampled areas.

DYS393 (Q-PLEX), the other contained DYS19, DYS385, DYS388, DYS91, DYS392 (E-PLEX). PCR reactions were performed in a final volume of 25 μl and contained 1× AmpliTaq Buffer (Applied Biosystem), 6–40 nmol of each primer, 1.25 μM dNTPs (Boeringer) and 1.25 units of AmpliTaq Gold (Applied Biosystem). Cycling parameters were: Q-PLEX: pre-incubation for 10 min at 94 °C, followed by 30 cycles of 45 s at 94 °C, 45 s at 54 °C, 1 min at 72 °C, and then a final incubation step of 72 °C for 7 min. E-PLEX: pre-incubation for 10 min at 94 °C, followed by 5 cycles of 45 s at 94 °C, 45 s at 58 °C, 1 min at 72 °C, then 30 cycles of 30 s at 94 °C, 45 s at 54 °C, 1 min at 72 °C, and then a final incubation step of 72 °C for 7 min.

Amplified samples were run on a 373ABI or 310ABI instruments (Applied Biosystem) and alleles called against an allelic ladder. PowerPlexY system was analysed following supplier indication (Promega). Allele nomenclature as previously described [2,3].

4. Quality controls

The laboratory participates to the quality control initiatives of the International Forensic Y chromosome group, the European DNA Profiling (EDNAP) group and the GeFI (Italian speaking working party of the ISFG). Alleles were identified by the use of an internal allelic ladder. A subset of the samples was checked for consistency by PowerPlex Y system (Promega).

5. Data analysis

Standard statistics were calculated by Arlequin Software [4].

6. Results

A total of 210 genotypes were produced in this study. The total number of different haplotypes across the 9-STR haplotype was 192 (Table 1). The unique assets were 178. The numbers of haplotypes per area were 68 (SA), 54 (MA) and 75 (TL). Sixty-four, 49 and 71 unique types were found in SA, MA and TL, respectively. Total number of shared types across the areas was five: three haplotypes between MA and TL and one each between SA and MA and between SA and TL. No haplotype was shared across all the three areas. The overall heterozygosity was 0.994, with values of 0.987, 0.980 and 0.986 in SA, MA and TL, respectively.

The occurrence of one locus duplication was detected at DYS19 (Hpt 192). This duplication was confirmed by re-typing after PowerPlex Y amplification. To compare this new collection of Italian Y chromosome haplotypes to a database previously published by the GeFI [5], the locus DYS388 was excluded in the haplotype asset. A total of 186 haplotypes were then defined by the 8-STRs haplotype comparison. One hundred and thirty-five haplotypes of our collection (72.5% of the total) were not present in the previous database. So a large number of new types suggests that the Y chromosome diversity defined by our set of markers in Italy is only partially known and that a considerable sample-to-sample heterogeneity exists. Additional and extensive sampling would then be needed to obtain a more comprehensive description of the haplotype composition in the Peninsula. Comparisons were conducted also using available published data on neighbour European (Greeks, Albanians, Kosovo Albanians and Croatians) and North African (Moroccan Saharawasis, Arabs and Berbers) populations [6–10] as summarised in Table 2. Interestingly, both Albanian samples shared an higher percentage of haplotypes and individuals than other geographically close Mediterranean samples. The three most common types found in our Italian dataset were

Table 1
Codes as indicated in the text: SA, South Apulia; MA, Marche; TL, Tuscany and Latium

	DYS19	DYS388	DYS389 I	DYS389 II	DYS390	DYS391	DYS392	DYS393	DYS385	SA	MA	TL
Hpt 1	12	12	14	31	24	10	11	13	14–18	1		
Hpt 2	13	9	13	31	25	10	11	13	16–17	1		
Hpt 3	13	12	12	29	22	11	16	11	16–16			1
Hpt 4	13	12	12	29	23	10	11	13	17–18	1		
Hpt 5	13	12	12	31	24	10	10	13	16–17	1		
Hpt 6	13	12	13	29	22	10	13	13	13–15	1		
Hpt 7	13	12	13	29	22	10	15	13	14–16			1
Hpt 8	13	12	13	29	23	11	13	13	11–14		1	
Hpt 9	13	12	13	29	23	13	13	13	11–12			1
Hpt 10	13	12	13	29	24	11	13	13	11–14			1
Hpt 11	13	12	13	29	24	11	13	14	11–13			1
Hpt 12	13	12	13	29	25	10	11	13	15–19	1		
Hpt 13	13	12	13	30	22	10	13	13	12–13	1		
Hpt 14	13	12	13	30	24	10	11	12	15–17			1
Hpt 15	13	12	13	30	24	10	11	13	16–18			1
Hpt 16	13	12	13	30	24	10	11	13	16–19		1	
Hpt 17	13	12	13	30	24	10	11	13	17–18	2	1	
Hpt 18	13	12	13	30	25	10	11	13	16–18			1
Hpt 19	13	12	13	30	25	10	12	13	16–18			1
Hpt 20	13	12	13	31	24	10	11	13	16–18	1		
Hpt 21	13	12	13	32	23	10	11	13	16–18			1
Hpt 22	13	12	14	31	23	10	11	13	16–17	1		
Hpt 23	13	12	14	31	23	11	11	12	16–16	1		
Hpt 24	13	12	14	31	24	10	11	13	17–19	1		
Hpt 25	13	12	14	32	24	11	11	13	11–16			1
Hpt 26	13	12	14	32	24	11	11	13	12–16			1
Hpt 27	13	12	15	30	24	10	11	13	17–17	1		
Hpt 28	13	12	15	31	24	11	11	13	16–16			1
Hpt 29	13	13	14	30	23	10	13	15	14–15	1		
Hpt 30	13	14	12	28	22	10	11	14	13–13		1	
Hpt 31	13	17	12	28	23	10	11	12	13–16	1		
Hpt 32	14	12	12	27	23	10	14	12	10–13			1
Hpt 33	14	12	12	28	23	11	13	13	11–14		1	
Hpt 34	14	12	12	28	24	11	13	12	11–14			1
Hpt 35	14	12	12	28	24	11	13	13	11–14		1	
Hpt 36	14	12	12	29	22	10	11	14	12–14			1
Hpt 37	14	12	12	29	22	10	11	14	13–13			1
Hpt 38	14	12	12	29	24	10	11	13	16–18		1	
Hpt 39	14	12	13	28	24	10	13	13	11–14			1
Hpt 40	14	12	13	28	24	11	13	13	11–14		1	
Hpt 41	14	12	13	29	23	10	13	13	11–14	2		
Hpt 42	14	12	13	29	23	10	15	13	16–19			1
Hpt 43	14	12	13	29	23	11	13	13	9–14			1
Hpt 44	14	12	13	29	23	11	13	13	11–11			1
Hpt 45	14	12	13	29	23	11	13	13	11–14		2	
Hpt 46	14	12	13	29	23	11	13	13	11–15	2		
Hpt 47	14	12	13	29	23	11	14	13	11–14			1
Hpt 48	14	12	13	29	23	11	15	12	16–16	1		
Hpt 49	14	12	13	29	24	10	13	12	11–14	1		2
Hpt 50	14	12	13	29	24	10	13	13	11–14		1	
Hpt 51	14	12	13	29	24	10	13	13	11–15		1	
Hpt 52	14	12	13	29	24	10	14	12	12–14	2		
Hpt 53	14	12	13	29	24	11	11	12	11–14	1		
Hpt 54	14	12	13	29	24	11	13	12	11–14			1
Hpt 55	14	12	13	29	24	11	13	13	11–12			1
Hpt 56	14	12	13	29	24	11	13	13	11–14		2	2
Hpt 57	14	12	13	29	24	11	13	13	12–13			1

Table 1 (*Continued*)

	DYS19	DYS388	DYS389 I	DYS389 II	DYS390	DYS391	DYS392	DYS393	DYS385	SA	MA	TL
Hpt 58	14	12	13	29	24	11	13	13	12–15		1	
Hpt 59	14	12	13	29	24	11	14	13	11–11	1		
Hpt 60	14	12	13	29	25	10	13	13	10–17			1
Hpt 61	14	12	13	29	25	11	13	14	11–14			1
Hpt 62	14	12	13	30	22	10	11	13	13–15	1		
Hpt 63	14	12	13	30	22	11	11	13	13–14	1		
Hpt 64	14	12	13	30	22	12	13	13	14–14	1		
Hpt 65	14	12	13	30	24	10	11	13	17–19	1		
Hpt 66	14	12	13	30	24	10	11	13	18–18	1		
Hpt 67	14	12	13	30	24	10	11	13	18–19	1		
Hpt 68	14	12	13	30	24	10	13	12	11–14		1	
Hpt 69	14	12	13	30	24	11	13	12	12–14			1
Hpt 70	14	12	13	30	24	11	13	12	12–15		1	
Hpt 71	14	12	13	30	24	11	13	13	11–14	1		
Hpt 72	14	12	13	30	24	11	13	13	11–15		1	
Hpt 73	14	12	13	30	24	11	13	13	11–16		1	
Hpt 74	14	12	13	30	25	11	13	12	12–14	1		
Hpt 75	14	12	13	31	23	11	11	13	16–16			1
Hpt 76	14	12	14	29	25	11	13	13	12–14		1	
Hpt 77	14	12	14	30	22	10	11	14	13–14			1
Hpt 78	14	12	14	30	23	11	13	12	11–14			1
Hpt 79	14	12	14	30	24	10	13	13	11–15	1		
Hpt 80	14	12	14	30	24	10	14	12	10–13			1
Hpt 81	14	12	14	30	24	11	13	13	11–11			1
Hpt 82	14	12	14	30	24	11	13	13	11–15		1	
Hpt 83	14	12	14	30	24	11	13	13	12–14	1		
Hpt 84	14	12	14	30	24	11	13	14	11–14			1
Hpt 85	14	12	14	31	24	10	11	12	16–18	1		
Hpt 86	14	12	15	30	22	10	11	13	13–14	1		
Hpt 87	14	13	13	29	23	10	11	12	13–19		1	
Hpt 88	14	13	13	29	24	10	13	13	11–14	1		
Hpt 89	14	13	13	29	24	11	13	13	11–14			1
Hpt 90	14	14	12	28	22	10	11	13	13–14			2
Hpt 91	14	14	12	28	23	10	11	13	13–14		1	
Hpt 92	14	14	13	29	22	10	11	13	13–15	1		
Hpt 93	14	15	13	29	22	10	11	12	13–14			2
Hpt 94	14	15	13	29	22	10	11	12	14–16		1	
Hpt 95	14	15	13	29	23	10	11	12	12–17		1	
Hpt 96	14	15	13	29	23	10	11	12	13–20	1		
Hpt 97	14	15	13	29	23	10	11	12	14–18			1
Hpt 98	14	15	13	29	23	10	11	13	13–17		1	
Hpt 99	14	15	13	29	23	11	11	12	13–20	1		
Hpt 100	14	15	13	30	22	10	11	12	14–16	1		
Hpt 101	14	15	13	30	22	10	11	13	13–16	1		
Hpt 102	14	15	13	30	23	10	11	12	13–17		1	
Hpt 103	14	15	13	30	23	10	11	13	13–19		1	
Hpt 104	14	15	14	29	23	8	11	12	12–12	1		
Hpt 105	14	15	14	29	26	10	11	13	14–15			1
Hpt 106	14	15	14	30	23	10	11	12	13–14	1		
Hpt 107	14	16	13	29	23	10	11	12	12–17		2	
Hpt 108	14	16	13	29	23	10	11	12	13–16		1	1
Hpt 109	14	16	13	29	23	10	11	12	14–17		2	
Hpt 110	14	16	13	30	22	10	11	12	13–14		1	
Hpt 111	14	16	13	30	23	10	11	12	13–18	1		
Hpt 112	14	16	13	30	23	10	11	12	13–19		1	
Hpt 113	14	16	14	30	24	9	11	12	13–14	1		
Hpt 114	14	17	13	29	23	10	11	12	12–17		1	
Hpt 115	14	17	13	29	23	10	11	12	12–17		1	
Hpt 116	14	17	14	32	24	10	11	10	13–17		1	

Table 1 (*Continued*)

	DYS19	DYS388	DYS389 I	DYS389 II	DYS390	DYS391	DYS392	DYS393	DYS385	SA	MA	TL
Hpt 117	14	18	13	29	23	10	11	12	13–17		1	
Hpt 118	14	18	13	29	25	10	11	12	15–17		1	
Hpt 119	14	18	14	30	23	10	11	12	13–17		1	
Hpt 120	15	11	13	29	22	10	11	14	14–15			1
Hpt 121	15	12	12	27	24	11	14	12	11–14			1
Hpt 122	15	12	12	28	21	10	11	13	12–14		1	
Hpt 123	15	12	12	28	22	9	11	14	12–14	1		
Hpt 124	15	12	12	28	23	10	11	14	15–15	1		
Hpt 125	15	12	12	28	23	10	12	14	14–15	1		
Hpt 126	15	12	12	28	23	10	12	14	15–15	1		
Hpt 127	15	12	12	28	24	11	11	12	15–17	1		
Hpt 128	15	12	12	29	21	10	11	15	14–15			1
Hpt 129	15	12	12	29	22	10	11	14	13–13	1		
Hpt 130	15	12	12	29	22	10	11	14	14–15		1	1
Hpt 131	15	12	12	30	23	10	11	14	13–14		1	
Hpt 132	15	12	13	28	23	11	13	13	11–14			1
Hpt 133	15	12	13	28	23	11	13	13	12–14		1	
Hpt 134	15	12	13	28	25	11	13	12	11–14			1
Hpt 135	15	12	13	29	24	11	13	13	11–15		1	
Hpt 136	15	12	13	29	25	11	13	12	14–14	1		
Hpt 137	15	12	13	29	25	11	13	13	11–14	1		
Hpt 138	15	12	13	29	26	10	13	14	11–13			1
Hpt 139	15	12	13	30	24	11	13	13	11–15			1
Hpt 140	15	12	13	30	25	11	13	12	11–14			1
Hpt 141	15	12	13	31	24	12	13	13	11–16	1		
Hpt 142	15	12	13	31	25	10	14	13	11–14			1
Hpt 143	15	12	14	30	24	10	13	13	13–14		1	
Hpt 144	15	12	14	31	21	11	11	13	14–16			1
Hpt 145	15	12	14	32	24	11	11	13	11–14			1
Hpt 146	15	12	16	30	23	10	13	14	12–14			1
Hpt 147	15	13	12	29	24	12	11	13	12–17		1	
Hpt 148	15	13	12	30	22	10	11	14	14–14			1
Hpt 149	15	13	13	27	24	11	12	13	16–16			1
Hpt 150	15	13	13	29	22	10	11	13	14–15			1
Hpt 151	15	13	13	29	22	10	11	14	14–15			1
Hpt 152	15	13	13	29	24	11	13	13	12–13		2	
Hpt 153	15	13	14	31	23	10	12	15	15–15	1		
Hpt 154	15	13	15	31	23	10	11	11	12–12		1	
Hpt 155	15	14	13	29	22	10	11	13	12–14			1
Hpt 156	15	14	14	32	24	10	11	12	14–14		1	
Hpt 157	15	15	12	28	23	10	11	12	14–17			1
Hpt 158	15	15	12	28	23	11	11	14	14–17	1		
Hpt 159	15	15	12	28	24	10	11	12	11–14			1
Hpt 160	15	15	12	28	24	10	11	12	15–17		1	
Hpt 161	15	15	12	28	24	10	11	12	16–18			1
Hpt 162	15	15	12	28	24	10	11	12	17–19	1		
Hpt 163	15	15	12	28	24	10	11	12	18–18	1		
Hpt 164	15	15	12	28	25	10	11	12	10–17	1		
Hpt 165	15	15	12	28	25	11	11	12	14–15			1
Hpt 166	15	15	12	29	22	9	11	12	13–16	1		
Hpt 167	15	15	12	29	24	10	11	12	14–17			1
Hpt 168	15	15	12	30	25	10	11	12	14–14		1	
Hpt 169	15	15	13	28	23	10	11	12	10–16			1
Hpt 170	15	15	13	28	24	10	11	12	13–18			1
Hpt 171	15	15	13	30	23	10	11	12	13–20	1		
Hpt 172	15	16	12	28	25	10	11	12	13–16			1
Hpt 173	15	16	12	29	23	9	11	12	13–14	1		
Hpt 174	15	16	13	29	23	10	11	11	12–15		1	
Hpt 175	15	16	13	30	23	9	11	13	13–16	1		

Table 1 (*Continued*)

	DYS19	DYS388	DYS389 I	DYS389 II	DYS390	DYS391	DYS392	DYS393	DYS385	SA	MA	TL
Hpt 176	15	16	14	30	23	10	11	12	13–16		1	
Hpt 177	16	12	12	27	24	10	12	14	14–14			1
Hpt 178	16	12	12	28	23	10	12	14	14–15	1		
Hpt 179	16	12	12	29	21	10	11	14	14–17		1	
Hpt 180	16	12	12	29	23	10	12	13	14–14	1		
Hpt 181	16	12	13	29	23	10	13	13	11–14	1		
Hpt 182	16	12	13	30	25	10	11	13	11–15	1		
Hpt 183	16	12	13	32	25	11	11	13	11–15		1	
Hpt 184	16	12	14	31	25	10	11	13	11–14			1
Hpt 185	16	12	14	31	26	10	11	13	11–14			1
Hpt 186	16	13	13	30	21	10	11	14	12–13	1		
Hpt 187	16	13	13	32	24	11	11	13	15–15		1	
Hpt 188	16	15	12	28	24	10	12	12	13–16			1
Hpt 189	16	15	12	29	24	10	11	12	16–17			1
Hpt 190	17	12	12	29	26	11	11	13	11–14			1
Hpt 191	17	13	13	32	24	11	11	13	14–15	1		
Hpt 192	15–16	12	14	32	22	10	11	13	13–15			1
										72	59	79

Indicated are the Y chromosome haplotypes and their occurrences in the three populations.

Table 2
Number of haplotypes/individuals shared with the three Italian population

Population	Greece	Albanians	Kossovo Albanians	Croatians	Morocco (three groups)[a]
Shared hpts/individuals	4/4	7/19	5/18	15/49	0/0; 2/3; 3/3
Total hpts/individuals	68/69	69/101	69/117	241/457	11/29; 26/44; 40/60

Reported are also the total numbers of haplotypes/individuals for each group. Data for Greece, Albanians and Kossovo Albanians did not include DYS388 locus; Croatian and Moroccan data did not include DYS385.

[a] Moroccan populations: Saharawis, Arabs and North-Central plus Southern Berbers.

haplotype 56, 17 and 49, represented 4, 3 and 3 times, respectively. We evaluated their occurrence in the largest worldwide Y chromosome haplotype collection present at www.yhrd.org. Results are summarised in Table 3. As expected, the largest number of match was with the Eurasian metapopulation. The matches in Asian and African metapopulations could be due to recent and historical gene flow [11,12] and/or homoplasy following recurrent mutations.

AMOVA analysis revealed significant genetic variation across the three Italian areas (percentage of genetic variation across populations: 1.25, $p < 0.05$). No significant population differentiation was found by using Fisher's Exact Test. Significant variation was found when the three samples were grouped and tested against the GeFI database (AMOVA, percentage of genetic variation across populations: 0.24, $p < 0.05$). The two groups were significantly different when using the Fisher's Exact Test for population differentiation ($p < 0.05$).

This paper follows the guidelines for publication of population data requested by the journal [13].

Table 3
Occurrences of the three most common haplotypes found in this study within the YHRD metapopulations (Release 15)

	Metapopulations				Total
	Eurasian (Italian)	East Asian	African	Others	
Hpt 56	583 (43)	2	21	–	606
Hpt 17	33 (5)	–	1	–	34
Hpt 49	29 (4)	–	–	–	29

DYS388 locus was excluded as not reported in the on-line database.

Acknowledgements

We would like to thank all the DNA donors that made this work possible. This research has been supported by a grant PRIN-MIUR 2002 (N.2002063871).

References

[1] S.A. Miller, D.D. Dykes, H.F. Polesky, A simple salting out procedure for extracting DNA from human nucleated cells, Nucleic Acids Res. 16 (1988) 1215.

[2] A. Carracedo, A. Beckmann, A. Bengs, B. Brinkmann, A. Caglia, C. Capelli, P. Gill, L. Gusmao, C. Hagelberg, C. Hohoff, B. Hoste, A. Kihlgren, A. Kloosterman, B. Myhre Dupuy, N. Morling, G. O'Donnell, W. Parson, C. Phillips, M.

Pouwels, R. Scheithauer, H. Schmitter, P.M. Schneider, J. Schumm, I. Skitsa, B. Stradmann-Bellinghausen, M. Stuart, D. Syndercombe Court, C. Vide, Results of a collaborative study of the EDNAP group regarding the reproducibility and robustness of the Y-chromosome STRs DYS19, DYS389 I and II, DYS390 and DYS393 in a PCR pentaplex format, Forensic Sci. Int. 119 (2001) 28–41.

[3] M. Kayser, A. Caglia, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, C. Schmitt, P.M. Schneider, R. Szibor, J. Teifel-Greding, G. Weichhold, P. de Knijff, L. Roewer, Evaluation of Y-chromosomal STRs: a multicenter study, Int. J. Legal Med. 110 (1997) 125–133.

[4] S. Schneider, D. Roessli, L. Excoffier, ARLEQUIN Ver. 2.000: A software for Population Genetic Data Analysis Genetics and Biometry Laboratory, University of Geneva, Geneva, 2000.

[5] S. Presciuttini, A. Caglia, M. Alu, A. Asmundo, L. Buscemi, L. Caenazzo, E. Carnevali, E. Carra, Z. De Battisti, F. De Stefano, R. Domenici, A. Piccinini, N. Resta, U. Ricci, V.L. Pascali, Y-chromosome haplotypes in Italy: the GEFI collaborative database, Forensic Sci. Int. 122 (2001) 184–188.

[6] K.S. Parreira, M.V. Lareu, P. Sanchez-Diz, I. Skitsa, A. Carracedo, DNA typing of short tandem repeat loci on Y-chromosome of Greek population, Forensic Sci. Int. 126 (2002) 261–264.

[7] C. Robino, S. Gino, U. Ricci, P. Grignani, C. Previdere, C. Torre, Y-chromosomal STR haplotypes in an Albanian population sample, Forensic Sci. Int. 129 (2002) 128–130.

[8] M. Pericic, L.B. Lauc, I.M. Klaric, B. Janicijevic, I. Behluli, P. Rudan, Y chromosome haplotypes in Albanian population from Kosovo, Forensic Sci. Int. 146 (2004) 61–64.

[9] L. Barac, M. Pericic, I.M. Klaric, B. Janicijevic, J. Parik, S. Rootsi, P. Rudan, Y chromosome STRs in Croatians, Forensic Sci. Int. 138 (2003) 127–133.

[10] E. Bosch, F. Calafell, A. Perez-Lezaun, D. Comas, H. Izaabel, O. Akhayat, A. Sefiani, G. Hariti, J.M. Dugoujon, J. Bertranpetit, Y chromosome STR haplotypes in four populations from northwest Africa, Int. J. Legal Med. 114 (2000) 36–40.

[11] E. Bosch, F. Calafell, D. Comas, P.J. Oefner, P.A. Underhill, J. Bertranpetit, High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula, Am. J. Hum. Genet. 68 (2001) 1019–1029.

[12] M.E. Hurles, C. Irven, J. Nicholson, P.G. Taylor, F.R. Santos, J. Loughlin, M.A. Jobling, B.C. Sykes, European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA, Am. J. Hum. Genet. 63 (1998) 1793–1806.

[13] P. Lincoln, A. Carracedo, A publication of population data of human polymorphisms, Forensic Sci. Int. 110 (2000) 3–5.

Forensic Science International: Genetics xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Short communication

Patterns of Y-STR variation in Italy

F. Brisighelli [a,b,e], A. Blanco-Verea [a], I. Boschi [b], P. Garagnani [c,d], V.L. Pascali [b], A. Carracedo [a,f], C. Capelli [d,e], A. Salas [a,*]

[a] Unidade de Xenética, Facultade de Medicina, Instituto de Ciencias Forenses, Universidade de Santiago de Compostela, Galicia, Spain
[b] Forensic Genetics Laboratory, Institute of Legal Medicine, Universita 'Cattolica del Sacro Cuore, Rome, Italy
[c] Dipartimento di Biologia Evoluzionistica Sperimentale, Universita 'di Bologna, Bologna, Italy
[d] Dipartimento di Patologia Sperimentale, Universitá di Bologna, Bologna, Italy
[e] Department of Zoology, University of Oxford, Oxford, UK
[f] Fundación Pública Galega de Medicina Xenómica (FPGMX-SERGAS), CIBER enfermedades raras, Santiago de Compostela, Galicia, Spain

ARTICLE INFO

Article history:
Received 8 November 2011
Received in revised form 19 February 2012
Accepted 11 March 2012

Keywords:
Italy
Haplotype
Yfiler
Forensics
Database

ABSTRACT

The 17 Y-chromosomal short tandem repeats (STRs) included in the AmpFlSTR Yfiler Amplification Kit (AB Applied Biosystems) (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385ab, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635 and GATA H4.1) were typed in 292 samples from seven Italian regions. Population comparisons with other European samples were undertaken; for this purpose, two databases were collated from the literature: (a) 19 population samples including >2900 Yfiler profiles, and (b) 67 population samples including >15,000 minimum haplotype profiles. A total of 276 different Yfiler haplotypes were observed in Italy, and only one of them was shared among our seven population samples. The overall haplotype diversity (0.9996) was comparable to other European samples. AMOVA indicates that among population variance depends on the amount of Y-STRs used, being higher when using minimal haplotypes. This is probably due to the fact that Yfiler profiles are represented by singleton haplotypes in all the population samples raising the diversity values to the maximum theoretical value. AMOVA results seems to depend even more strongly on the amount of population samples used, the among population variance in Italy ranging from 2.82% to 11.03% (using 15 and 32 Italian populations samples, respectively). Variance is not as strongly stratified geographically within Italy, although it is notorious that latitude is more important than longitude in the distribution of variance. The results also indicated that Italy is less stratified than other European samples. The present study contributes to enrich the Y-chromosome databases regarding high-resolution Y-chromosome data sets and demonstrates that extended Y-STR profiles substantially increases the discriminatory capacity in individual identification for forensic purposes.

1. Introduction

The identification of differences within Italian national borders is of special interest for the understanding of human population history and setting up forensic reference databases. A significant amount of genetic analysis has been conducted in Europe but the coverage is not homogenous across the area. Within this context, the data on the Italian Peninsula is scanty, with only few studies focusing on a limited number of samples and few genetic systems. Despite this, the available Y chromosome data have allowed to outline a broad indication of within-country population differences, with migration and admixture events being the major responsible of the current distribution of genetic variation. A major cline across the Italian Peninsula has been described [1], but local drift and founder effect have been signaled as the main explanation for the observed distribution of genetic diversity. The North to South Y chromosome clines support an admixture model of the Mesolithic original inhabitants of the peninsula and the incoming Neolithic farmers from the Middle East [2].

Several studies have been undertaken in various Italian regions, most of them focusing on the analysis of the minimum haplotype defined by seven Y-STRs [2–11]. Only few studies were carried out using the Yfiler STR commercial kit that allows genotyping 17 Y-STRs [12–16]. Analysis of this extended Y-STR datasets has demonstrated to be very useful in forensic genetics [17–19] because they increase the discrimination power very substantially when compared to the figures reached using minimum haplotypes.

Here we have genotyped the largest sample set of Italian samples analyzed to date using the Yfiler with the main aim of contributing to improve Y-chromosome forensic databases.

* Corresponding author at: Unidade de Xenética, Departamento de Anatomía Patológica e Ciencias Forenses, Universidade de Santiago de Compostela, Galicia, Spain. Tel.: +34 981 582327; fax: +34 981 580336.
E-mail address: antonio.salas@usc.es (A. Salas).

1872-4973/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2012.03.003

Table 1
Diversity indices computed for different Italian regions on the Y-STRs. Population codes are as indicated in Section 2. The label "All Italy" refers to the full set of population samples analyzed in the present study pooled in a single group, while the label "Europe" refers to a pooled group made from the European population sets listed in Table S1.

Population	Region	Ref.	*n*	*K*	*DC*	*UH*	*HD*	*h*	*M*
CLI	NW	p.s.	45	40	0.89	38	0.9919 ± 0.0079	0.6453 ± 0.3345	9.680 ± 4.5005
EFR	NE	p.s	47	43	0.91	42	0.9954 ± 0.0059	0.6420 ± 0.3325	9.630 ± 4.4941
CMA	CE	p.s.	38	37	0.97	36	0.9986 ± 0.0065	0.6213 ± 0.3258	8.698 ± 4.1058
WCA	SW	p.s.	27	27	1	27	1.0000 ± 0.0101	0.6484 ± 0.3413	9.726 ± 4.5980
CCA	SE	p.s.	31	30	0.97	29	0.9978 ± 0.0089	0.6528 ± 0.3433	9.139 ± 4.3210
SIC	SW	p.s.	57	56	0.98	56	0.9994 ± 0.0034	0.6434 ± 0.3335	9.008 ± 4.2096
SAP	SE	p.s.	47	44	0.94	41	0.9972 ± 0.0051	0.6396 ± 0.3314	9.594 ± 4.4784
All Italy	–	p.s.	292	276	0.94	276	0.9996 ± 0.0004	0.6553 ± 0.3328	9.830 ± 4.5127
Europe	–	*	1928	1846	0.96	1557	1.0000 ± 0.0101	0.6362 ± 0.3228	9.544 ± 4.3790

Population and region codes are as in Fig. 1. *K* = number of different haplotypes; *n* = sample size; *DC* = discrimination capacity (defined as the number of different haplotypes divided by the sample size); *UH* = number of unique haplotypes; *HD* = haplotype diversity; *h* = gene diversity over loci; *M* = average number of pairwise differences. Standard deviations are reported for *HD*, *h* and *M*, meaning the values for both sampling and stochastic processes
* [26,27,29,34,39,41,58–61].

Several AMOVA were also undertaken in order to evaluate the level of population stratification within the country and in comparison to other European populations, both using minimum and extended Y-choromosome haplotypes.

2. Material and methods

2.1. Samples

A total of 292 samples were analyzed from seven different Italian areas, namely, Central Liguria ($n = 45$; CLI), East Friuli ($n = 47$; EFR), Central Marche ($n = 38$; CMA), West Calabria ($n = 27$; WCA), Central Campania ($n = 31$; CCA), Sicily ($n = 57$; SIC), and South Apulia ($n = 47$; SAP). The geographic location and the number of samples for each locality are shown in Table 1. Individuals were (closely) paternally unrelated, they had different surnames and they defined themselves as belonging to a paternal lineage residing in the area from at least three generations.

2.2. DNA extraction

Collection was performed by buccal swab and blood drawing after informed consent. Blood extraction was undertaken following a salting-out method [20], modified and readapted to buccal cells.

2.3. PCR

The AmpFlSTR Yfiler Amplification Kit (AB, Applied Biosystem) (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385ab, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635 and GATA H4.1) was genotyped in all the samples. PCR amplification was performed according to the manufacturer's conditions (AmpFlSTR Yfiler PCR Amplification Kit, AB Applied Biosystem).

2.4. Y-STR genotyping

PCR products were analyzed by capillary electrophoresis in an ABI 3100 Genetic analyzer (Applied Biosystem, Foster City, CA). Allele assignment was carried out by comparison with reference sequenced ladders [21]. Alleles for GATA H4.1 locus were named according to the ISFG recommendations [22] by adding nine repeats to the nomenclature included in the typing kit [21]. The DYS389II allele number was determined by subtracting the DYS389I repeat number. In our investigation a number of Y-STR duplications were found. Haplotypes containing duplicated loci were not included in analysis based on STR variation [2,23].

2.5. Quality control

The population data obtained in the present study were submitted to the Y chromosome Haplotype Reference Database under the YHRD accession number: from YA003721 to YA003726 and YA003744. All the analyses were carried out at the laboratory of the University of Santiago de Compostela; the laboratory participates in the annual proficiency testing of the GHEP-ISFG WG (http://www.gep-isfg.org).

2.6. Statistical analysis

Allele and haplotype frequencies were estimated by haplotype counting. Haplotype diversity (*HD*; sometimes referred to as the gene diversity) was calculated for each population sample using the equation $HD = (1 - \sum q_i^2)(n/n - 1)$, where n is the sample size and q_i is the haplotype frequency. Gene diversity (h) is equivalent to *HD* but referred to single loci; it can be computed also as an average over loci. The discriminatory capacity (*DC*) was determined by dividing the number of different haplotypes by the number of samples in a given population [24].

Arlequin software v.3.5.1.2 [25] was used to calculate population pairwise genetic distances (R_{ST}) and perform analysis of molecular variance (AMOVA) tests. In order to examine the relationship between the Italian and other neighboring populations, R_{ST} genetic distances were used to generate multi-dimensional scaling (*MDS*) plots using the software Statistica 7 (http://www.statsoft.com). *MDS* was undertaken using additional data from other European populations [2,4–6,8,11,13,15,16,26–54]; the minimal haplotype recommended by the YHRD but excluding the system DYS385ab (DYS19, DYS389 I, DYS389 II, DYS390, DYS391, DYS392, DYS393) [55,56] was used. For some analysis, population samples were clustered into main national regions, considering North Italy the region located North of the Po river, South the region located South of Latina city, and Central, the populations located in between North and South Italy. In addition, we consider the Apennines to be a natural barrier separating East and West Italy (see Fig. 1).

References and geographic information of the populations samples used in the present study to undertake inter-population comparisons (*e.g.* AMOVA, MDS) are given in Table S1.

3. Results

3.1. Genetic diversity

All the Italian samples were genotyped for 17 Y-chromosome linked STRs. The total number of different haplotypes observed was 276 (among 292 samples); 263 of them were unique in our sample set. In the whole Italian sample, eleven haplotypes

Please cite this article in press as: F. Brisighelli, et al., Patterns of Y-STR variation in Italy, Forensic Sci. Int. Genet. (2012), doi:10.1016/j.fsigen.2012.03.003

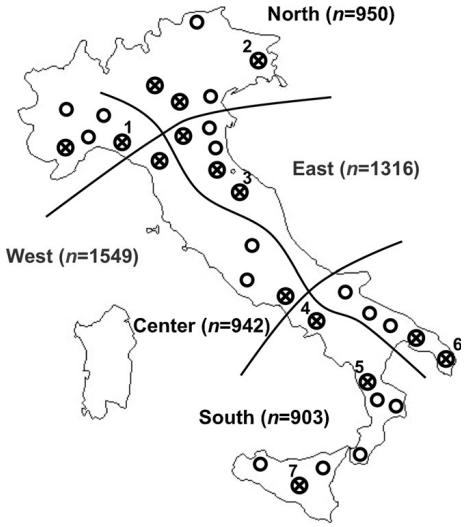


Fig. 1. Map of Italy, showing the location of the samples analyzed in the present study plus other Italian samples collected from the literature. Population codes for the populations analyzed in the present study: (1) = CLI: Central Liguria (n = 45); (2) = EFR: East Friuli Venezia-Giulia (n = 47); (3) = CMA: Central Marche (n = 38); (4) = CCA: Central Campania (n = 31); (5) = WCA: West Calabria (n = 27); (6) = SAP: South Apulia (n = 47); (7) = SIC: Sicily (n = 47). Crossed dots indicate those samples that were genotyped for the YFiler, while open dots are the samples that were genotyped only for the minimal haplotype.

appeared twice, and just two haplotypes appeared three and four times, respectively. Allele frequencies and *GD* for each Y-STR loci analyzed in the population samples, haplotype numbers and the most frequent haplotypes within the samples are shown in Tables S2 and S3, respectively. The highest locus diversity (h) was observed at DYS456 (0.795) while the lowest value was found at DYS391 (0.488). Gene diversity over loci was quite similar across populations (0.66 on average). Values of *GD* and *DC* for each of the seven different population samples are shown in Table 1.

West Calabria is the sample that accounts for the highest values of molecular diversity (with the exception of h, where Central Campania reaches an slightly higher value) among the seven samples analyzed in the present study. One individual from Central Liguria and one from South Apulia with a duplicate allele (15, 16 and 17) at DYS19 were observed. This duplication has been previously described by other authors [2,55–57] and has been associated with the Y-chromosome haplogroups C3c or G. Four intermediate allelic variants (12.2, 17.2, 18.2 and 19.2) at DYS458 locus were found in thirteen samples: one in Central Liguria, three in Central Marche, two in Central Campania, five in Sicily and two in South Apulia. One individual with an allelic variant at DYS448 (18.2) was observed.

Considering the complete Yfiler haplotype (17 Y-STRs), our seven population samples shared only one haplotype. An average *HD* of 0.9996 ± 0.0004 and an overall *DC* of 0.948 were obtained for the seven populations.

When considering the minimal haplotype (seven Y-STRs), 27 profiles were shared between the seven populations. Also for the minimal haplotype, in Central Liguria and Central Marche, East Friuli, West Calabria, Central Campania, South Apulia and Sicily, 32, 38, 25, 26, 34 and 44 different haplotypes were observed, respectively. All the seven populations yielded similar haplotype diversities. The R_{ST} pairwise distance values indicated significant differences between Central Liguria and East Friuli, Central Marche and Sicily, and between East Friuli and West Calabria, Sicily, Central Marche and South Apulia (data not shown). The most common haplotype was 14–13–16–24–11–13–13 with frequencies of 15.5%, 7.4% 6.4%, 3.5% and 2.2%, in Central Liguria, West Calabria, Central Campania, Sicily and South Apulia populations, respectively.

3.2. AMOVA analysis

A total of 2,990 Yfiler (17 Y-STR) and 15,032 minimal haplotype (7 Y-STRs) profiles were collected from the literature [2,4,5,15,16,26,27,29,38,39,41,58–74] (see Table S1).

Table 2
Y chromosome AMOVA in different Italian and European populations. For the analysis carried out within Italy, we considered the samples analyzed in the present study plus other samples collected from the literature. In order to investigate the effect of sampling in Europe, AMOVA was also carried out eliminating a different population sample at the time as indicated in the table. When Yfiler data was available, the analyses were performed using the full Yfiler set (column 17-YSTRs) and also reducing these profiles to their minimal haplotypes (column 7 Y-STRs); however, note that some populations were only available for the minimal haplotype (last three rows in the table). For the analysis carried out in the European context, all the Italian samples were pooled. References for population samples are given in the main text and supplementary data. The column "Pop" indicates the number of population samples, while "n" refers to the total sample size. Statistical significance was assessed using a permutation procedure, as implemented in Arlequin (10,000 permutations); all p-values were below 0.01.

	Pop	n	17 Y-STR		7 Y-STRs	
			Among	Within	Among	Within
17-YSTRs available						
Italy (present study)	7	292	1.91	98.09	2.42	97.58
Italy (literature)	8	766	1.85	98.15	2.33	97.67
Italy (all)	15	1058	2.34	97.66	2.82	97.18
Europe (all)	23	3282	7.71	92.29	12.81	87.19
Europe minus Croatia	22	2916	6.99	93.01	10.61	89.39
Europe minus Serbia	22	3097	8.00	92.00	13.29	86.71
Europe minus Romania	22	3160	7.99	92.10	13.40	86.60
Europe minus Poland	22	3027	6.37	93.63	10.61	89.39
Europe minus Portugal	22	2651	7.47	92.53	12.90	87.10
Europe minus Catalonia	22	3234	7.48	92.52	12.85	87.15
Europe minus Greece	22	3091	7.82	92.18	13.60	86.40
Europe minus Austria	22	3152	8.16	91.84	13.54	86.46
Europe minus Italy	22	2224	9.95	90.05	15.16	84.84
Only 7 YSTRs available						
Italy (literature)	25	2770	–	–	11.55	88.45
Italy (all)	32	3062	–	–	11.03	88.97
Europe (all)	52	15032	–	–	10.78	89.22

Please cite this article in press as: F. Brisighelli, et al., Patterns of Y-STR variation in Italy, Forensic Sci. Int. Genet. (2012), doi:10.1016/j.fsigen.2012.03.003

AMOVA was performed following different grouping schemes in order to investigate the distribution of variance in Italy but also in an European context. As shown in Table 2, there were not notable differences in the among population variance observed in Italy when using Yfiler and minimum haplotypes; however, it is paradoxical that the values were always slightly higher when using minimum haplotypes. This is probably due to the fact that using Yfiler profiles increases the amount of singleton haplotypes in all the sample populations, then raising their haplotype diversity to the maximum theoretical value (see Table 1). The differences in among population variance when using Yfiler *versus* minimal haplotype are more marked when analyzing other European samples (see below).

The AMOVA analysis performed in this study also shows that it is not only the amount of genetic information available (Yfiler *versus* minimal haplotype) that determines the results of AMOVA, but even more important seems to be the amount of populations samples used. For instance, AMOVA of Italian samples, leads to among population variance to rank from 2.82% when using 15 samples to 11.03% when using 32 population samples (Table 2) for the minimum haplotype. The higher latter figure compares well with the among population variance observed in Europe (that ranks from 10.61% to 15.16%), also for the minimal haplotype. In order to disregard the influence that population outliers (*e.g.* generated by genetic drift or artificially by way of genotyping errors) could have in this analysis, we carried out an MDS analysis based on R_{ST} distances; this analysis does not show distintive patterns among the 32 population samples available for the minimum haplotype or the 15 samples available for the Yfiler (Fig. S1).

Eliminating Italy from the full set of European samples increases the among population variance to the highest value (9.95% using 17 Y-STRs and 15.16% using 7 Y-STRs), while eliminating Poland from the European AMOVA lead among population variance to its minimum (6.37% using 17 Y-STRs and 10.61% using 7 Y-STRs); Table 2. This fact indirectly indicates that Italy is less stratified that other countries; in other words, the presence of Italy dilutes among population variance in Europe.

AMOVA was also undertaken in population samples distributed by main geographical regions (Table 3). The results indicated that geography have limited impact on variance apportionment; for instance, within population variance is always higher than 96% in every population scenario, independently of the number of populations considered or the amount of genetic information used (Yfiler *versus* minimum haplotypes) (Table 3). The analysis showed however that latitude has a slightly higher impact than longitude; thus, among groups variance ranged between 1% and 2% when comparing North, Central and South, while it is always 0% when comparing East and West Italy.

Table 3
Y chromosome AMOVA in different Italian geographical regions ($p < 0.01$). As in Table 2 analyses were carried for the Yfiler and the minimal haplotype. References for population samples are given in the main text and supplementary data. Statistical significance was assessed using a permutation procedure, as implemented in Arlequin (10,000 permutations); all *p*-values were below 0.01.

	North/Center/South	West/East
Present study [7 pop]		
17 Y-STRs		
Among groups	2.58	0
Within pops	97.28	98.43
Among pops within groups	0.14	1.57
7 Y-STRs		
Among groups	1.94	0
Within pops	96.97	98.18
Among pops within groups	1.09	1.82
Present study + literature [15 pop]		
17 Y-STRs		
Among groups	1.18	0
Within pops	97.32	97.85
Among pops within groups	1.50	2.15
7 Y-STRs		
Among groups	1.28	0
Within pops	96.80	97.45
Among pops within groups	1.92	2.55

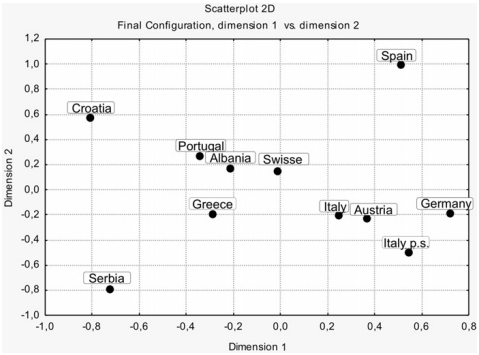


Fig. 2. MDS of Italian and other neighboring European samples. Stress for 2D MDS = 0.26; mean stress if considering the stress values for MDS-scaled randomly generated matrices of fifteen objects (800 matrices per dimension) [79].

3.3. Multi dimensional scaling

To visualize the existing genetic relationships between population samples we used *MDS* based on the matrix of pairwise R_{ST} values. The *MDS* plot shows (Fig. 2) for the dimension 1 and 2, a close relationship between our Italian sample to Austria, Germany and other Italian populations collected from the literature (as confirmed from the R_{ST} genetic distances matrix). Spain is also closely related to Italy in the dimension 1, but clearly separates from Italy in the dimension 2.

4. Discussion

The goal of the present study was to contribute to an Italian database of forensic interest given that: (i) most of the Y chromosome data available in the literature is restricted in the number of loci (minimal haplotype), (ii) generally focus on single populations, or (iii) cover the Italian peninsula only partially. Our sampling comprises individuals from North, Central and South of Italy and has been genotyped for the full set of Y-STRs considered in the Yfiler. We have shown that the analysis of full Y-STR profiles (Yfiler) *versus* minimal haplotypes is important for the detection of local genetic differences, and that the genetic discrimination provided for the Yfiler profiles is substantiality larger than for the minimal haplotype.

We have performed AMOVA analysis in different population contexts, and we have demonstrated that measure of population stratification strongly depends not only on the amount of genetic information used, but also more importantly, on the amount of population samples employed in the analysis [75]. This guarantees more caution when using F_{ST} values to correct for population stratification in forensic casework given that this estimates strongly depend on the sampling scheme. Similar problems apply to other countries and population groups; e.g. [76,77].

Although Italy seems to show lower levels of among population variance than other European regions, these values are high

Please cite this article in press as: F. Brisighelli, et al., Patterns of Y-STR variation in Italy, Forensic Sci. Int. Genet. (2012), doi:10.1016/j.fsigen.2012.03.003

enough (about 3–11%; Table 2) to guaranty caution when using the Y-chromosome test in forensics. Thus, the results raise the issue of developing local reference databases in Italy for forensic purposes given that substantial differences existing across the Italian Peninsula. The present study provides additional information on the genetic variation of the Italian population and the necessity to enlarge the number of Y STRs typed in order to better understand the substructure of the Italian Peninsula. This paper follows the guidelines for publication of population data requested by the journal [78].

Acknowledgments

We thank all persons and local communities that donated their DNA and made the present study feasible. We additionally thank all the people that were involved in the sampling: Anna Barbaro, Fidelia Cascini, Donata Luiselli, Sara Partemi, Angela Reveruzzi, Daniela Vantaggiato. This study received support from the Ministerio de Ciencia e Innovación (SAF2008-02971 and SAF2011-26983) (A.S.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2012.03.003.

References

[1] F. Di Giacomo, F. Luca, N. Anagnou, G. Ciavarella, R.M. Corbo, M. Cresta, F. Cucci, L. Di Stasi, V. Agostiano, M. Giparaki, A. Loutradis, C. Mammi', E.N. Michalodimitrakis, F. Papola, G. Pedicini, E. Plata, L. Terrenato, S. Tofanelli, P. Malaspina, A. Novelletto, Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects, Mol. Phylogenet. Evol. 28 (3) (2003) 387–395.
[2] C. Capelli, F. Brisighelli, F. Scarnicci, B. Arredi, A. Caglià, G. Vetrugno, S. Tofanelli, V. Onofri, A. Tagliabracci, G. Paoli, V.L. Pascali, Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter, Mol. Phylogenet. Evol. 44 (1) (2007) 228–239.
[3] A. Barbaro, P. Cormaci, G. Falcone, M. Rizzo, Genetic study of 11 Y-STRs in the populations of Reggio Calabria, Catanzaro, Cosenza (Calabria—South of Italy), Forensic Sci. Int. 146 (Suppl.) (2004) S129–S131.
[4] N. Cerri, A. Verzeletti, B. Bandera, F. De Ferrari, Population data for 12 Y-chromosome STRs in a sample from Brescia (northern Italy), Forensic Sci. Int. 152 (1) (2005) 83–87.
[5] G. Ferri, S. Ceccardi, F. Lugaresi, C. Bini, F. Ingravallo, A. Cicognani, M. Falconi, S. Pelotti, Male haplotypes and haplogroups differences between urban (Rimini) and rural area (Valmarecchia) in Romagna region (North Italy), Forensic Sci. Int. 175 (2–3) (2008) 250–255.
[6] S. Presciuttini, A. Caglià, M. Alù, A. Asmundo, L. Buscemi, L. Caenazzo, E. Carnevali, E. Carra, Z. De Battisti, F. De Stefano, R. Domenici, A. Piccinini, N. Resta, U. Ricci, V.L. Pascali, Y-chromosome haplotypes in Italy: the GEFI collaborative database, Forensic Sci. Int. 122 (2–3) (2001) 184–188.
[7] C. Rapone, A. Geraci, C. Capelli, A. De Meo, G. D'Errico, F. Barni, A. Berti, G. Lago, Y chromosome haplotypes in central-south Italy: implication for reference database, Forensic Sci. Int. 172 (1) (2007) 67–71.
[8] C. Robino, S. Inturri, S. Gino, C. Torre, C. Di Gaetano, F. Crobu, V. Romano, G. Matullo, A. Piazza, Y-chromosomal STR haplotypes in Sicily, Forensic Sci. Int. 159 (2–3) (2006) 235–240.
[9] N. Cerutti, A. Marin, C. Di Gaetano, P. Pappi, F. Crobu, F. Riccardino, G. Matullo, A. Piazza, Population data for Y-chromosome STR haplotypes from Piedmont (Italy), Forensic Sci. Int. 158 (2–3) (2006) 238–243.
[10] M.E. Ghiani, I.S. Piras, R.J. Mitchell, G. Vona, Y-chromosome 10 locus short tandem repeat haplotypes in a population sample from Sicily Italy, Leg. Med. 6 (2004) 89–96.
[11] C. Capelli, V. Onofri, F. Brisighelli, I. Boschi, F. Scarnicci, M. Masullo, G. Ferri, S. Tofanelli, A. Tagliabracci, L. Gusmão, A. Amorim, F. Gatto, M. Kirin, D. Merlitti, M. Brión, A.B. Verea, V. Romano, F. Calì, V. Pascali, Moors and Saracens in Europe: estimating the medieval North African male legacy in southern Europe, Eur. J. Hum. Genet. 17 (6) (2009) 848–852.
[12] I. Carboni, A.L. Nutini, B. Porfirio, M. Genuardi, U. Ricci, Genetic STRs variation in a large population from Tuscany (Italy), Forensic Sci. Int. Genet. 1 (3–4) (2007) e10–e11.
[13] G. Ferri, M. Alù, B. Corradini, E. Radheshi, G. Beduschi, Slow and fast evolving markers typing in Modena males (North Italy), Forensic Sci. Int. Genet. 3 (2) (2009) e31–e33.
[14] V. Onofri, F. Alessandrini, C. Turchi, B. Fraternale, L. Buscemi, M. Pesaresi, A. Tagliabracci, Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis, Int. J. Legal Med. 121 (3) (2007) 234–237.
[15] S. Turrina, R. Atzei, D. De Leo, Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17 plex loci PCR assay, Int. J. Legal Med. 120 (1) (2006) 56–59.
[16] S. Pelotti, C. Bini, A. Barbaro, L. Caenazzo, E. Carnevali, N. Cerri, R. Dominici, G. Ferri, M. Maniscalco, V. Onofri, A. Piccinini, C. Previderè, U. Ricci, C. Robino, F. Scarnicci, F. Torricelli, M. Venturi, S. Presciuttini, G.s.g.o.Y.-c. characterization, Microgeographic variation of Y-chromosome haplotypes in Italy, Forensic Sci Int: Genet. Suppl. Ser. 1 (2008) 239–241.
[17] E.K. Hanson, J. Ballantyne, An ultra-high discrimination Y chromosome short tandem repeat multiplex DNA typing system, PLoS One 2 (8) (2007) e688.
[18] A.E. Decker, M.C. Kline, P.M. Vallone, J.M. Butler, The impact of additional Y-STR loci on resolving common haplotypes and closely related individuals, Forensic Sci. Int. Genet. 1 (2) (2007) 215–217.
[19] C. Alves, L. Gusmão, J. Barbosa, A. Amorim, Evaluating the informative power of Y-STRs: a comparative study using European and new African haplotype data, Forensic Sci. Int. 134 (2–3) (2003) 126–133.
[20] S.A. Miller, D.D. Dykes, H.F. Polesky, A simple salting out procedure for extracting DNA from human nucleated cells, Nucleic Acids Res. 16 (3) (1988) 1215.
[21] J.J. Mulero, C.W. Chang, L.M. Calandro, R.L. Green, Y. Li, C.L. Johnson, L.K. Hennessy, Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system, J. Forensic Sci. 51 (1) (2006) 64–75.
[22] L. Gusmão, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, W.R. Mayr, N. Morling, M. Prinz, L. Roewer, C. Tyler-Smith, P.M. Schneider, DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis, Int. J. Legal Med. 120 (4) (2006) 191–200.
[23] C. Capelli, F. Brisighelli, F. Scarnicci, A. Blanco-Verea, M. Brión, V.L. Pascali, Phylogenetic evidence for multiple independent duplication events at the DYS19 locus, Forensic Sci. Int. Genet. 1 (3–4) (2007) 287–290.
[24] M. Kayser, A. Caglià, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, C. Schmitt, L. Roewer, et al., Evaluation of Y-chromosomal STRs: a multicenter study, Int. J. Legal Med. 110 (3) (1997), 125–133, 141–129.
[25] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (3) (2010) 564–567.
[26] M.L. Pontes, L. Caine, D. Abrantes, G. Lima, M.F. Pinheiro, Allele frequencies and population data for 17 Y-STR loci (AmpFlSTR Y-filer) in a Northern Portuguese population sample, Forensic Sci. Int. 170 (1) (2007) 62–67.
[27] J. Ljubković, A. Stipišić, D. Šutlović, M. Definis-Gojanović, K. Bučan, S. Andelinović, Y-chromosomal short tandem repeat haplotypes in southern Croatian male population defined by 17 loci, Croat. Med. J. 49 (2) (2008) 201–206.
[28] E. Rossi, B. Rolf, M. Schürenkamp, B. Brinkmann, Y-Chromosome STR haplotypes in an Italian population sample, Int. J. Legal Med. 112 (1998) 78–81.
[29] N. Haliti, M. Carapina, M. Masić, D. Strinović, I.M. Klarić, M. Kubat, Evaluation of population variation at 17 autosomal STR and 16 Y-STR haplotype loci in Croatians, Forensic Sci. Int. Genet. 3 (4) (2009) e137–e138.
[30] P. Martín, J. García-Hirschfeld, O. García, L. Gusmão, P. García, C. Albarrán, M. Sancho, A. Alonso, A Spanish population study of 17 Y-chromosome STR loci, Forensic Sci. Int. 139 (2–3) (2004) 231–235.
[31] L. Lovrečić, S. Ristić, B. Brajenović, M. Kapović, B. Peterlin, Human Y-specific STR haplotypes in the Western Croatian population sample, Forensic Sci. Int. 149 (2–3) (2005) 257–261.
[32] M. Carvalho, M.J. Anjos, L. Andrade, V. Lopes, M.V. Santos, J.J. Gamero, F. Corte Real, M.C. Vide, Y-chromosome STR haplotypes in two population samples: Azores Islands and Central Portugal, Forensic Sci. Int. 134 (1) (2003) 29–35.
[33] V. Prieto, Y. Torres, M.J. Farfán, M. López-Soto, J. García-Hirschfeld, P. Sanz, Population genetics of Y chromosomal STR haplotypes in south Spain (Andalusia), Int. Congr. Ser. 1239 (2003) 403–408.
[34] C. Sánchez, C. Barrot, A. Xifró, M. Ortega, I.G. de Aranda, E. Huguet, J. Corbella, M. Gené, Haplotype frequencies of 16 Y-chromosome STR loci in the Barcelona metropolitan area population using Y-Filer kit, Forensic Sci. Int. 172 (2–3) (2007) 211–217.
[35] C. Robino, S. Gino, U. Ricci, P. Grignani, C. Previderè, C. Torre, Y-chromosomal STR haplotypes in an Albanian population sample, Forensic Sci. Int. 129 (2) (2002) 128–130.
[36] C. Robino, S. Varacalli, S. Gino, A. Chatzikyriakidou, A. Kouvatsi, C. Triantaphyllidis, C. Di Gaetano, F. Crobu, G. Matullo, A. Piazza, C. Torre, Y-chromosomal STR haplotypes in a population sample from continental Greece, and the islands of Crete and Chios, Forensic Sci. Int. 145 (1) (2004) 61–64.
[37] C. Haas, T. Wangensteen, N. Giezendanner, A. Kratzer, W. Bär, Y-chromosome STR haplotypes in a population sample from Switzerland (Zurich area), Forensic Sci. Int. 158 (2–3) (2006) 213–218.
[38] L. Kovatsi, J.L. Saunier, J.A. Irwin, Population genetics of Y-chromosome STRs in a population of Northern Greeks, Forensic Sci. Int. Genet. 4 (1) (2009) e21–e22.
[39] I.S. Veselinović, D.M. Zgonjanin, M.P. Maletin, O. Stojković, M. Djurendić-Brenesel, R.M. Vuković, M.M. Tasić, Allele frequencies and population data for 17 Y-chromosome STR loci in a Serbian population sample from Vojvodina province, Forensic Sci. Int. 176 (2–3) (2008) e23–e28.
[40] L.B. Lauč, M. Peričić, I.M. Klarić, A. Šijački, D. Popović, B. Janićijević, P. Rudan, Y chromosome STR polymorphisms in a Serbian population sample, Forensic Sci. Int. 150 (1) (2005) 97–101.

[41] B. Berger, A. Lindinger, H. Niederstätter, P. Grubwieser, W. Parson, Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay, Int. J. Legal Med. 119 (4) (2005) 241–246.
[42] H. Rodig, M. Grum, H.D. Grimmecke, Population study and evaluation of 20 Y-chromosome STR loci in Germans, Int. J. Legal Med. 121 (1) (2007) 24–27.
[43] C. Hohoff, K. Dewa, U. Sibbing, K. Hoppe, P. Forster, B. Brinkmann, Y-chromosomal microsatellite mutation rates in a population sample from northwestern Germany, Int. J. Legal Med. 121 (5) (2007) 359–363.
[44] U.D. Immel, M. Kleiber, M. Klintschar, Y chromosome polymorphisms and haplotypes in South Saxony-Anhalt (Germany), Forensic Sci. Int. 155 (2–3) (2005) 211–215.
[45] U. Schmidt, N. Meier, S. Lutz, Y-chromosomal STR haplotypes in a population sample from southwest Germany (Freiburg area), Int. J. Legal Med. 117 (4) (2003) 211–217.
[46] A.I. Zurita, A. Hernández, J.J. Sánchez, J.A. Cuellas, Y-chromosome STR haplotypes in the Canary Islands population (Spain), Forensic Sci. Int. 148 (2–3) (2005) 233–238.
[47] J.A. Pena, S. García-Obregón, A.M. Pérez-Miranda, M.M. De Pancorbo, M.A. Alfonso-Sánchez, Gene flow in the Iberian Peninsula determined from Y-chromosome STR loci, Am. J. Hum. Biol. 18 (4) (2006) 532–539.
[48] A.M. López, S. Álvarez, L. Gusmão, C. Alves, M.S. Mesa, A. Albentosa, G. Arribas, R. López, P.A. Barrio, A. Amorim, E. Arroyo-Pardo, Population data for 16 Y-chromosome STRs in four populations from Pyrenees (Spain), Forensic Sci. Int. 140 (1) (2004) 125–129.
[49] M. Gené, N. Borrego, A. Xifró, E. Piqué, P. Moreno, E. Huguet, Haplotype frequencies of eight Y-chromosome STR loci in Barcelona (North-East Spain), Int. J. Legal Med. 112 (6) (1999) 403–405.
[50] O. García, P. Martín, L. Gusmão, C. Albarrán, S. Alonso, C. de la Rua, C. Flores, N. Izagirre, R. Penas, J. Antonio Perez, I. Uriarte, I. Yurrebaso, A. Alonso, A Basque country autochthonous population study of 11 Y-chromosome STR loci, Forensic Sci. Int. 145 (1) (2004) 65–68.
[51] M.T. Zarrabeitia, J.A. Riancho, P. Sánchez-Diz, P. Sánchez-Velasco, 7-Locus Y chromosome haplotype profiling in a northern Spain population, Forensic Sci. Int. 123 (1) (2001) 78–80.
[52] M. Nata, B. Brinkmann, B. Rolf, Y-chromosomal STR haplotypes in a population from north west Germany, Int. J. Legal Med. 112 (1999) 406–408.
[53] G. Jiménez, A. Picornell, C. Tomàs, J.A. Castro, M.M. Ramón, Y-chromosome polymorphism data in Majorcan, Minorcan and Valencian populations (eastern Spain), Forensic Sci. Int. 124 (2001) 231–234.
[54] V. Rodríguez, C. Tomàs, J.J. Sánchez, J.A. Castro, M.M. Ramon, A. Barbaro, N. Morling, A. Picornell, Genetic sub-structure in western Mediterranean populations revealed by 12 Y-chromosome STR loci, Int. J. Legal Med. 123 (2) (2009) 137–141.
[55] M. Kayser, P. de Knijff, P. Dieltjes, M. Krawczak, M. Nagy, T. Zerjal, A. Pandya, C. Tyler-Smith, L. Roewer, Applications of microsatellite-based Y chromosome haplotyping, Electrophoresis 18 (9) (1997) 1602–1607.
[56] S. Willuweit, L. Roewer, Y chromosome haplotype reference database (YHRD): update, Forensic Sci. Int. Genet. 1 (2) (2007) 83–87.
[57] P. Balaresque, E.J. Parkin, L. Roewer, D.R. Carvalho-Silva, R.J. Mitchell, R.A. van Oorschot, J. Henke, M. Stoneking, I. Nasidze, J. Wetton, P. de Knijff, C. Tyler-Smith, M.A. Jobling, Genomic complexity of the Y-STR DYS19: inversions, deletions and founder lineages carrying duplications, Int. J. Legal Med. 123 (1) (2009) 15–23.
[58] C. Alves, V. Gomes, M.J. Prata, A. Amorim, L. Gusmão, Population data for Y-chromosome haplotypes defined by 17 STRs (AmpFlSTR YFiler) in Portugal, Forensic Sci. Int. 171 (2–3) (2007) 250–255.
[59] I. Soltyszewski, W. Pepinski, M. Spolnicka, E. Kartasinska, M. Konarzewska, J. Janica, Y-chromosomal haplotypes for the AmpFlSTR Yfiler PCR amplification kit in a population sample from Central Poland, Forensic Sci. Int. 168 (1) (2007) 61–67.
[60] F. Stanciu, V. Cuţăr, S. Pîrlea, V. Stoian, I.M. Stoian, O. Sevastre, O.R. Popescu, Population data for Y-chromosome haplotypes defined by 17 STRs in South-East Romania, Leg. Med. (Tokyo) 12 (5) (2010) 259–264.
[61] A.M. Bento, M. Carvalho, V. Lopes, A. Serra, H.A. Costa, L. Andrade, F. Balsa, C. Oliveira, L. Batista, J. Gamero, M.J. Anjos, L. Gusmão, F. Corte-Real, Distribution of Y-chromosomal haplotypes in the central Portuguese population using 17-STRs, Forensic Sci. Int. Genet. 4 (1) (2009) e35–e36.
[62] R. Pawlowski, A. Dettlaff-Kakol, Population data of nine Y-chromosomal STR loci in northern Poland, Forensic Sci. Int. 131 (2–3) (2003) 209–213.
[63] L.E. Barbarii, B. Rolf, D. Dermengiu, Y-chromosomal STR haplotypes in a Romanian population sample, Int. J. Legal Med. 117 (5) (2003) 312–315.
[64] D.J. Ballard, C. Phillips, C.R. Thacker, D.S. Court, Y chromosome STR haplotype data for an Irish population, Forensic Sci. Int. 161 (1) (2006) 64–68.
[65] D.J. Ballard, C. Phillips, C.R. Thacker, C. Robson, A.P. Revoir, D. Syndercombe Court, Y chromosome STR haplotypes in three UK populations, Forensic Sci. Int. 152 (2–3) (2005) 289–305.
[66] K. De Maesschalck, E. Vanhoutte, K. Knaepen, N. Vanderheyden, J.J. Cassiman, R. Decorte, Y-chromosomal STR haplotypes in a Belgian population sample and identification of a micro-variant with a flanking site mutation at DYS19, Forensic Sci. Int. 152 (1) (2005) 89–94.
[67] B.M. Dupuy, M. Stenersen, T.T. Lu, B. Olaisen, Geographical heterogeneity of Y-chromosomal lineages in Norway, Forensic Sci. Int. 164 (1) (2006) 10–19.
[68] E. Ehler, R. Marvan, D. Vanek, Evaluation of 14 Y-chromosomal short tandem repeat haplotype with focus on DYS449, DYS456, and DYS458: Czech population sample, Croat. Med. J. 51 (1) (2010) 54–60.
[69] J. Zastera, L. Roewer, S. Willuweit, P. Sekerka, L. Benesova, M. Minarik, Assembly of a large Y-STR haplotype database for the Czech population and investigation of its substructure, Forensic Sci. Int. Genet. 4 (3) (2010) e75–e78.
[70] C. Hallemberg, K. Nielsen, B. Simonsen, J. Sánchez, N. Morling, Y-chromosome STR haplotypes in Danes, Forensic Sci. Int. 155 (2005) 205–210.
[71] G. Holmlund, H. Nilsson, A. Karlsson, B. Lindblom, Y-chromosome STR haplotypes in Sweden, Forensic Sci. Int. 160 (1) (2006) 66–79.
[72] M. Spiroski, T. Arsov, C. Kruger, S. Willuweit, L. Roewer, Y-chromosomal STR haplotypes in Macedonian population samples, Forensic Sci. Int. 148 (1) (2005) 69–73.
[73] B. Zaharova, S. Andonova, A. Gilissen, J.J. Cassiman, R. Decorte, I. Kremensky, Y-chromosomal STR haplotypes in three major population groups in Bulgaria, Forensic Sci. Int. 124 (2–3) (2001) 182–186.
[74] S. Furedi, J. Woller, Z. Padar, M. Angyal, Y-STR haplotyping in two Hungarian populations, Int. J. Legal Med. 113 (1) (1999) 38–42.
[75] P.G. Meirmans, P.W. Hedrick, Assessing population structure: F(ST) and related measures, Mol. Ecol. Resour. 11 (1) (2010) 5–18.
[76] U. Toscanini, L. Gusmão, G. Berardi, A. Amorim, Á. Carracedo, A. Salas, E. Raimondi, Testing for genetic structure in different urban Argentinian populations, Forensic Sci. Int. 165 (1) (2007) 35–40.
[77] U. Toscanini, L. Gusmão, G. Berardi, A. Amorim, Á. Carracedo, A. Salas, E. Raimondi, Y chromosome microsatellite genetic variation in two Native American populations from Argentina: Population stratification and mutation data, Forensic Sci. Int. Genet. 2 (4) (2008) 274–280.
[78] A. Carracedo, J.M. Butler, L. Gusmão, W. Parson, L. Roewer, P.M. Schneider, Publication of population data for forensic purposes, Forensic Sci. Int. Genet. 4 (3) (2010) 145–147.
[79] K. Sturrock, J. Rocha, A multidimensional scaling stress evaluation table, Field Methods 12 (1) (2000) 49–60.

Please cite this article in press as: F. Brisighelli, et al., Patterns of Y-STR variation in Italy, Forensic Sci. Int. Genet. (2012), doi:10.1016/j.fsigen.2012.03.003

Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics 1 (2007) 287–290
FSI GENETICS
www.elsevier.com/locate/fsig

Short communication

Phylogenetic evidence for multiple independent duplication events at the DYS19 locus

Cristian Capelli [a,*], Francesca Brisighelli [b,c], Francesca Scarnicci [b], Alejandro Blanco-Verea [c], Maria Brion [c], Vince L. Pascali [b]

[a] *Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, OX1 3PS Oxford, UK*
[b] *Istituto di Medicina Legale e delle Assicurazioni, Universita' Cattolica del S. Cuore, Rome, Italy*
[c] *Grupo de Medicina Xenómica, Instituto de Medicina Legal, Facultad de Medicina, Universidad de Santiago de Compostela, Galicia, Spain*

Received 2 February 2007; received in revised form 10 May 2007; accepted 1 June 2007

Abstract

Duplication events at Y chromosome STR loci have been repeatedly described in human populations. DYS19 is probably the best known example and it exhibits duplicate state in individuals from all continents. Despite the large amount of available data, evolutionary relationship between DYS19 duplication-bearing chromosomes has not been so far investigated. We address the genealogical correlation among such chromosomes by analysing newly identified DYS19 duplicated Y chromosomes by SNP genotyping and microsatellite-based network analysis. SNP and network analysis show that DYS19 duplicated Y chromosomes associate with different Y chromosome lineages. These results indicate that DYS19 duplication occurred more than once during human evolution.

Keywords: DYS19; Duplication; Y chromosome microsatellites; Network analysis; Y chromosome genealogy

1. Introduction

Y chromosome STRs have today a steady place in several fields of molecular analysis. In population genetics and in forensics, a repertory of these Y-linked markers is routinely used to help identify sexual assaulters, analyze genealogical relationships and reconstructing ancient population migrations. The community of Y chromosome analysis users has been consequently – and since the beginning – made familiar with the typical haploid pattern of these markers, almost invariably reproducing the PCR analysis scheme of 'one locus-one band/peak'.

On the other hand, locus duplications placed along some highly redundant Y chromosome portions [1] can exist and create a PCR-doublets phenomenon. For example, DYS389 and DYS385 STR loci are duplicated in both humans and chimpanzees [2]. Recently, other locus duplication examples have been described at Y chromosome loci ([3], and reference therein).

This phenomenon is interesting from a population and evolutionary point of view. When duplication of a given STR locus is seen across the generality (or the vast majority) of human populations, the obvious question arises as to whether it occurred once or by independent reiteration of the same event. Noticing haploid doublets has implications for the forensic genetics practitioner too [3].

In view of this series of considerations, a better understanding of the evolutionary history of Y chromosome duplication is essential. We here try to tackle this issue by focusing on the well know example of DYS19 locus duplication, whose abundance of observational data may supply the ideal subject for phylogenetic analysis.

2. Materials and methods

A collection of newly identified DYS19 duplicated chromosomes were available to us, framed within a high-resolution male haplotypic set of data ([4,15]; Table 1). All duplicated chromosomes have been confirmed by amplification using the PowerPlex Y system ([5]; see Table 1). Haplogroups assignment was done by scoring P15, P16 and M286 markers using the following primers: P15-F-ccaatgcttgaggttctgaat,

* Corresponding author at: Tel.: +44 1865 271261; fax: +44 1865 310447.
E-mail address: cristian.capelli@zoo.ox.ac.uk (C. Capelli).

1872-4973/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2007.06.001

Table 1
Y chromosome haplotypes bearing DYS19 duplications found in Italy, all within the G2*(xG2a, b) lineage

n	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385	DYS437	DYS438	DYS439
2	13–14	14	16	22	10	11	14	13-13	16	10	12
1	13–14	15	16	22	10	11	14	13-13	17	10	12
1+ 1[a]	14–16	14	16	22	11	11	14	13-14	16	10	12
1[b]	15–16	14	18	22	10	11	13	13-15	16	10	12
1[a]	15–16	14	18	22	10	11	13	12-15	16	10	12
1[a]	15–17	12	17	22	11	11	13	14-14	16	10	11
1[a]	15–17	14	16	22	11	11	13	14-16	16	10	12
1	15–17	13	19	24	10	11	15	14-15	15	10	9

Number of repeats at DYS389II locus was considered after subtracting DYS389I repeats.
[a] Haplotype described in Capelli et al. [15].
[b] Haplotype described in Capelli et al. [4].

P15-R-agagcctcaatccatcatgc; P16-F-agcacacagttgagcaatgg, P16-R-ccggcaacagatcagaaact; M286-F-cgttgtccctgtgtttccat, M286-R-ggattgcaggcatcagcta. The three markers identify G2, G2a and G2b haplogroups, respectively. Phylogenetic network was calculated by Network 4.112 (fluxus-engineering.com [7]) applying the Median Joining algorithms sequentially. A total of 133 DYS19 duplicated Y chromosome were collected from available literature (88 from the www.yhrd.com—release 20; 35 from Zerjal et al. [8]; 10 from Capelli et al. [4,15]). The duplicated haplotypes described by Nasidze et al. [9] were already included in the YHRD. Of these, 96 genotyped for the extended Y-STR haplotype loci plus DYS438 and DYS439, were available and were included in the network analysis. DYS19 and DYS385 were excluded due to difficulties in assigning each of the duplicated allele to the corresponding locus, and DYS392 resulted monomorphic for allele 11. The following loci were used for network calculation: DYS389I, DYS389II (subtracting DYS389I repeats), DYS390, DYS391, DYS393, DYS438, DYS439. One haplotype from the Y-STR repository (YHRD) showing a tri-allelic DYS19 pattern was included in the analysis. A weighting scheme was applied giving to each locus a score as follows: $[V_A \times 10]/V_L$, considering V_A as the average variance of the repeat score estimated across loci and V_L the variance at the repeat score at each specific locus.

3. Results and discussion

It is well known that DYS19 locus PCR doublets exist worldwide in humans (see www.yhrd.org). In the context of a SNP genotyping study, Zerjal et al. [8] showed that Asian individuals bearing this locus duplication belong to C3c haplogroup (Y chromosome genealogy [10]) (identified by marker M48, Chris Tyler-Smith, personal communication). More recently, Nasidze et al. [9] described DYS19 duplications in populations from the Caucasus and they classified them within haplogroup C3c. In the course of a large Italian Y chromosome screening, we have now identified 10 additional DYS19 duplication-bearing chromosomes—with the relevant frequency of occurrence amounting to 1% ([4,15, unpublished data]). By additional typing work, we can report that all these chromosomes classify in the haplogroup G2*(xG2a,b)—a branch of its own in the Y chromosome genealogical hierarchy, identified by P15, P16 and M286 markers. Our finding clearly implies that duplication at DYS19 had to occur at least twice, driven by as many *independent* genetic events.

Starting from this line of evidence, we collected all DYS19 duplication-bearing haplotypes available in the literature and used the inherent specific archive of data to generate a microsatellite network. The resulting structure (Fig. 1) plots in two clusters. The first cluster mostly accommodates chromosomes of European origin, including the Italian chromosomes that bear the P15 mutation. The second cluster is made mostly of Russian and Kazakh individuals carrying the M48 mutation. The most common DYS19 doublets were 15–16 in group 1 and 16–17 in group 2. The European cluster appears to be more heterogeneous than the second one, as suggested by the number of multiple mutational events connecting the various haplotypes and a larger microsatellite variance (0.45 versus 0.12). This could reflect partial knowledge (by ineffective sampling) of the whole set of G2 DYS19 duplicated chromosomes—and potentially also additional, independent, still uncharacterised duplication events.

In the light of the current existence of the two lineages, the single-origin hypothesis implies that the original duplication event had to occur very early in the timescale of human evolution [10]. However, the currently available world-wide set of genotyped samples from different Y chromosome genealogy branches shows a limited number of DYS19 duplication-bearing chromosomes, suggesting that, in the case of single origin, multiple reversion events should have occurred. The multiple-origin hypothesis, requesting so far a minimum of only two independent duplication events, appears, as a more parsimonious alternative, the most likely.

We note that the DYS19 locus lies within the chromosomal region delimited by the IR3/IR3 repeats that have experienced several molecular rearrangements [11]. While reporting on the structural variation of a larger overlapping area, Jobling et al. [12] recently noticed a number of independent deletion events to be generated by non-allelic homologous recombination. According to these authors, a parallel event of reciprocal duplication should have occur at the same time—as it is suggested by the existence of two chromosomes showing the duplicated state at the DYS458 locus mapping within the deleted/potentially duplicated region. The same mechanism could account of the DYS19 duplication, by a larger

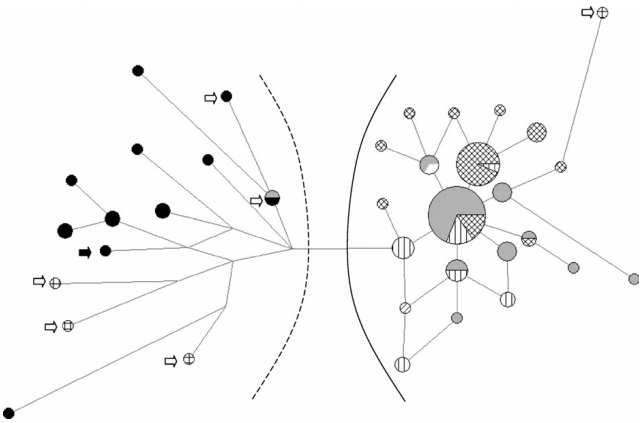


Fig. 1. Phylogenetic network. Total number of individuals in the network is 96. The size of the pies is proportional to the number of individuals represented, being one the minimum number of chromosome per indicated node. The different individuals' origins are coded within the pies as follows: grey, Russia; diagonal lines, Kazakhstan; vertical lines, Asia excluding Russia and Kazakhstan; white, Africa; black, Europe; crossed lines, African American, South American and Hispanics. G2*(xG2a,G2b) and C3c haplotypes are defined by the dashed and the solid lines, respectively. Arrows point to nodes containing no SNP defined haplotypes. For nomenclature consistency, three repeats have been added to the DYS389I alleles from Zerjal et al. [8] dataset (Chris Tyler-Smith, personal communication).

recombination event involving the entire IR3/IR3 region. Gene conversion could as well be an alternative mechanism. Several regions homologous to DYS19 have been identified within the Y chromosome [13], among which the IR3 regions bear probably the highest homology (99.75% [1]). Regions as these encompass up to 30% of the Y chromosome euchromatin [1] and have experienced multiple gene conversion events [14]. Therefore, some of them are ideal candidates as counterparts in the process of DYS19 duplication.

The status of neighbouring markers should be of decisive help to identify the exact mechanism of production (duplication-deletion versus gene conversion). Further molecular characterisation of DYS19 duplication-bearing chromosomes will help to clarify this issue.

Acknowledgements

Authors would like to thank all DNA donors who made this work possible. CC additionally would like to acknowledge Chris Tyler-Smith and Lutz Roewer for kindly providing data. This study was partially funded by the Italian Ministry of University (PRIN-MIUR 2002, N.2002063871).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2007.06.001.

References

[1] H. Skaletsky, T. Kuroda-Kawaguchi, P.J. Minx, H.S. Cordum, L. Hillier, L.G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S.F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.P. Yang, R.H. Waterston, R.K. Wilson, S. Rozen, D.C. Page, The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes, Nature 423 (2003) 825–837.

[2] L. Gusmao, A. Gonzalez-Neira, C. Alves, P. Sanchez-Diz, E.M. Dauber, A. Amorim, A. Carracedo, Genetic diversity of Y-specific STRs in chimpanzees (Pan troglodytes), Am. J. Primatol. 57 (2002) 21–29.

[3] J.M. Butler, A.E. Decker, M.C. Kline, P.M. Vallone, Chromosomal duplications along the Y-chromosome and their potential impact on Y-STR interpretation, J. Forensic Sci. 50 (2005) 853–859.

[4] C. Capelli, B. Arredi, L. Baldassari, I. Boschi, F. Brisighelli, A. Caglia, M. Dobosz, F. Scarnicci, G. Vetrugno, V.L. Pascali, A 9-loci Y chromosome haplotype in three Italian populations, Forensic Sci. Int. 159 (2006) 64–70.

[5] B.E. Krenke, L. Viculis, M.L. Richard, M. Prinz, S.C. Milne, C. Ladd, A.M. Gross, T. Gornall, J.R. Frappier, A.J. Eisenberg, C. Barna, X.G. Aranda, M.S. Adamowicz, B. Budowle, Validation of male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex, Forensic Sci. Int. 151 (2005) 111–124.

[7] H.-J. Bandelt, P. Forster, A. Röhl, Median-joining networks for inferring intraspecific phylogenies, Mol. Biol. Evol. 16 (1999) 37–48.

[8] T. Zerjal, Y. Xue, G. Bertorelle, R.S. Wells, W. Bao, S. Zhu, R. Qamar, Q. Ayub, A. Mohyuddin, S. Fu, P. Li, N. Yuldasheva, R. Ruzibakiev, J. Xu, Q. Shu, R. Du, H. Yang, M.E. Hurles, E. Robinson, T. Gerelsaikhan, B. Dashnyam, S.Q. Mehdi, C. Tyler-Smith, The genetic legacy of the Mongols, Am. J. Hum. Genet. 72 (2003) 717–721.

[9] I. Nasidze, D. Quinque, I. Dupanloup, R. Cordaux, L. Kokshunova, M. Stoneking, Genetic evidence for the Mongolian ancestry of Kalmyks, Am. J. Phys. Anthropol. 128 (2005) 846–854.

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

[11] S. Repping, S.K. van Daalen, L.G. Brown, C.M. Korver, J. Lange, J.D. Marszalek, T. Pyntikova, F. van der Veen, H. Skaletsky, D.C. Page, S. Rozen, High mutation rates have driven extensive structural polymorphism among human Y chromosomes, Nat. Genet. 8 (2006) 463–467.

[12] M.A. Jobling, I.C. Lo, D.J. Turner, G.R. Bowden, A.C. Lee, Y. Xue, D. Carvalho-Silva, M.E. Hurles, S.M. Adams, Y.M. Chang, T. Kraaijenbrink, J. Henke, G. Guanti, B. McKeown, R.A. van Oorschot, R.J. Mitchell, P. de

Knijff, C. Tyler-Smith, E.J. Parkin, Structural variation on the short arm of the human Y chromosome: recurrent multigene deletions encompassing Amelogenin Y, Hum. Mol. Genet. 16 (2007) 307–316.
[13] J.M. Butler, R. Schoske, Duplication of DYS19 flanking regions in other parts of the Y chromosome, Int. J. Legal Med. 118 (2004) 178–183.
[14] S. Rozen, H. Skaletsky, J.D. Marszalek, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, D.C. Page, Abundant gene conversion between arms of palindromes in human and ape Y chromosomes, Nature 423 (2003) 873–876.
[15] C. Capelli, F. Brisighelli, F. Scarnicci, B. Arredi, A. Caglia', G. Vetrugno, V. Onofri, S. Tofanelli, G. Paoli, V.L. Pascali, Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic–Neolithic encounter, Mol. Phyl. Evol. 44 (2007) 228–239.

Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics 3 (2009) e29–e30
FSI GENETICS
www.elsevier.com/locate/fsig

Announcement of Population Data

Allele frequencies of fifteen STRs in a representative sample of the Italian population

F. Brisighelli [a,b,*], C. Capelli [c], I. Boschi [b], P. Garagnani [d], M.V. Lareu [a], V.L. Pascali [b], A. Carracedo [a]

[a] *Institute of Legal Medicine, Genomics Medicine Group, University of Santiago de Compostela, Santiago de Compostela, Spain*
[b] *Forensic Genetics Laboratory, Institute of Legal Medicine, Università Cattolica del Sacro Cuore, Rome, Italy*
[c] *Department of Zoology, University of Oxford, Oxford, UK*
[d] *Dipartimento di Biologia evoluzionistica sperimentale, Università di Bologna, Bologna, Italy*

Received 12 February 2008; received in revised form 6 May 2008; accepted 7 May 2008

Abstract

Fifteen autosomal short tandem repeat (STR) markers (D3S1358, HUMTH01, D21S11, D18S51, Penta E, D5S818, D13S317, D7S820, D16S539, CSF1PO, Penta D, HUMvWA, D8S1179, HUMTPOX and FGA) were analyzed in more than 400 unrelated individuals from nine different areas of Italy. After Bonferroni correction, no evidence of population structure was identified, either by considering each population as independent or by combining populations according to their geographic origin (North, Central and South of Italy). Forensic indexes were estimated considering all samples together. Combined power of discrimination (PD) and combined power of exclusion (PE) for the 15 tested STR loci were 0.9999999997 and 0.964708775, respectively. Low genetic distances were found between our data and those previously published for other neighboring European populations.

Keywords: STRs; Population data; Powerplex 16; Italy

1. Population

A sample of 441 individuals from nine different areas of Italy were analyzed. These areas correspond to different Italian regions: Liguria, Friuli-Venezia Giulia, Marche, Latium, Calabria, Campania, Puglia and Sicily. Collection was performed by buccal swab and blood drawing after informed consent.

2. Extraction

DNA extraction was performed with a salting-out method [1].

3. PCR

PowerPlex16 multiplex system (Promega Corporation, Madison, WI) was amplified according to the manufacturer's conditions [2].

4. Typing

PCR products were analyzed by capillary electrophoresis in an ABI 3100 Genetic analyzer (Applied Biosystem, Foster City, CA). Allele assignment was carried out by comparison with reference sequenced ladders [2].

5. Analysis of data

Arlequin software ver 3.11 [3] was used to calculate allele frequencies, population pairwise genetic distances (F_{ST}), AMOVA test, and also to assess departures from Hardy–Weinberg equilibrium. Statistical parameters of forensic interest (Power of Discrimination, Power of Exclusion and Matching Probability) were calculated using PowerStats v1.2 (Promega Corporation, Madison, WI) software package.

6. Quality control

Proficiency testing of the GEP-ISFG WG (http://www.gep-isfg.org) and GEDNAP blind trials (http://www.gednap.de.vu).

* Corresponding author at: Institute of Legal Medicine, Genomics Medicine Group, University of Santiago de Compostela, Rúa San Francisco, s/n, 15782; Santiago de Compostela, A Coruña, España. Tel.: +34 981582327; fax: +34 981580336.
E-mail address: fbrisi@libero.it (F. Brisighelli).

1872-4973/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2008.05.002

7. Results

The whole genotype data set, allele frequencies and genetic distances between the studied sample and other European populations [4–10] are available as e-component.

8. Other remarks

Population differentiation test showed no significant differences between populations. The AMOVA test was performed considering each population independently or grouped according to their geographic origin (North, Central and South of Italy). Results showed that most of the genetic variation occurs within individuals, underlining no evidence of population structure (data not shown).

Deviation from Hardy–Weinberg equilibrium has been detected for the HUMTH01, for the D7S820, for D16S539, for the Penta D, and finally for the HUMvWA. Such deviations disappeared after Bonferroni correction.

The combined power of exclusion (PE) and power of discrimination (PD) for the fifteen studied loci were 0.964708775 and 0.9999999997, respectively. The combined matching probability value was 1 in 3.33×10^8. Based on heterozygosity and polymorphic information content (PIC), FGA may be considered as the most informative loci.

Locus-by-locus allelic frequencies were compared to previously published Italian and Mediterranean population data [4–10] (e-component). After applying the Bonferroni correction for multiple tests, population differentiation tests showed that Italy had significant differences with Bosnia–Herzegovina in 4 out of 13 loci (D5S818, D7S820, D16S539 and CSF1PO), with Spain in 3 out of 13 loci (D3S1358, TH01 and D16S539) and with Kosovo Albanians in 1 out of 13 loci (D18S51).

A number of Italian databases are currently available [11–14]. However, those are either restricted in the number of loci, focus on single populations or cover the Italian peninsula only partially. Our sample, by comprising individuals from North, Central and South of Italy, provides additional information on the genetic variation of the Italian population.

This paper follows the guidelines for publication of population data requested by the journal [15].

Acknowledgements

We would like to thank all the DNA donors that made this work possible. We additionally thank all the people that were involved in the sampling: Anna Barbaro, Fidelia Cascini, Donata Luiselli, Sara Partemi. We also would like to thank the AVIS Blood collection directors of Norma, Giuseppe Santucci and Sezze, Ubaldo Brandolini. The technical assistance of Amelia Rodríguez and Raquel Calvo is highly appreciated.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2008.05.002.

References

[1] S.A. Miller, D.D. Dykes, H.F. Polesky, A simple salting out procedure for extracting DNA from human nucleated cells, Nucleic Acids Res. 16 (1988) 1215.
[2] B.E. Krenke, A. Tereba, S.J. Anderson, E. Buel, S. Culhane, C.J. Finis, C.S. Tomsey, J.M. Zachetti, A. Masibay, D.R. Rabbach, E.A. Amiott, C.J. Sprecher, Validation of a 16-locus fluorescent multiplex system, J. For. Sci. 47 (2002) 773–785.
[3] L. Excoffier, G. Laval, S. Schneider, Arlequin ver. 3.0: an integrated software package for population genetics data analysis, Evol. Bioinform. (Online) 1 (2005) 47–50.
[4] P. Sánchez-Diz, P. G. Menounos, A. Carracedo, I. Skitsa, 16 STR data of a Greek population, Forensic Sci Int: Genetics doi:10.1016/j.fsigen. 2008.01.002 http://dx.doi.org/10.1016/j.fsigen.2008.01.002.
[5] S. Presciuttini, N. Cerri, S. Turrina, B. Pennato, M. Alù, A. Asmundo, A. Barbaro, I. Boschi, L. Buscemi, L. Caenazzo, E. Carnevali, D. De Leo, C. Di Nunno, R. Domenici, M. Maniscalco, G. Peloso, S. Pelotti, A. Piccinini, D. Podini, U. Ricci, C. Robino, L. Saravo, A. Verzeletti, M. Venturi, A. Tagliabracci, Validation of a large Italian Database of 15 STR loci, For. Sci. Int. 156 (2006) 266–268.
[6] M. Kubat, J. Skavić, I. Behluli, B. Nuraj, T. Bekteshi, M. Behluli, I. Martinović Klarić, M. Pericić, Population genetics of the 15 AmpF /STR Identifiler loci in Kosovo Albanians, Int. J. Legal Med. 118 (2004) 115–118.
[7] M.F. Pinheiro, L. Caine, L. Pontes, D. Abrantes, G. Lima, M.J. Pereira, P. Rezende, Allele frequencies of sixteen STRs in the population of Northern Portugal, For. Sci. Int. 148 (2005) 221–223.
[8] D. Marjanovic, N. Bakal, N. Pojskic, L. Kapur, K. Drobnic, D. Primorac, K. Bajrovic, R. Hadziselmovic, Allele frequencies for 15 short tandem repeat loci in a representative sample of Bosnians and Herzegovinians, For. Sci. Int. 156 (2006) 79–81.
[9] P. Projić, V. Skaro, I. Samija, N. Pojskić, A. Durmić-Pasić, L. Kovacević, N. Bakal, D. Primorac, D. Marjanović, Allele frequencies for 15 short tandem repeat loci in representative sample of Croatian population, Croat. Med. J. 48 (2007) 4730–4737.
[10] M.V. Camacho, C. Benito, A.M. Figueiras, Allelic frequencies of the 15 STR loci included in the AmpFlSTR Identifiler PCR Amplification Kit in an autochthonous sample from Spain, For. Sci. Int. 173 (2007) 241–245, Epub 2007 Mar 8.
[11] S. Presciuttini, F. Ciampini, M. Alù, N. Cerri, M. Dobosz, R. Domenici, G. Peloso, S. Pelotti, A. Piccinini, E. Ponzano, U. Ricci, A. Tagliabracci, J.E. Baley-Wilson, F. De Stefano, V. Pascali, Allele sharing in first-degree and unrelated pairs of individuals in the GeFI AmpFlSTR® Profiler Plus™ database, For. Sci. Int. 131 (2003) 85–89.
[12] L. Garofano, M. Pizzamiglio, C. Vecchio, G. Lago, T. Floris, G. D'Errico, G. Brembilla, A. Romano, B. Budowle, Italian population data on thirteen short tandem repeat loci: HUMTH01, D21S11, D18S51, HUMVWFA31, HUMFIBRA, D8S1179, HUMTPOX, HUMCSF1PO, D16S539, D7S820, D13S317, D5S818, D3S1358, For. Sci. Int. 97 (1998) 53–60.
[13] R. Biondo, A. Spinella, P. Montagna, P.S. Walsh, C. Holt, B. Budowle, Regional Italian Allele frequencies at nine short tandem repeat loci, For. Sci. Int. 115 (2001) 95–98.
[14] R. Maviglia, M. Dobosz, I. Boschi, A. Caglià, D. Hall, C. Capelli, E. d'Aloja, M. Pescarmona, A. Moscetti, V.L. Pascali, G. Destro-Bisol, A repository of 14 PCR-loci Italian gene frequencies in the World Wide Web, For. Sci. Int. 115 (2001) 99–101.
[15] P. Lincoln, A. Carracedo, Publication of population data of human polymorphisms, For. Sci. Int. 110 (2000) 3–5.

Forensic Science International: Genetics 5 (2011) 548–549

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Letter to the Editor

Allele frequencies of the new European Standard Set (ESS) loci in the Italian population

Dear Editor,

Allele frequencies of five new STR loci (D22S1045, D10S1248, D1S1656, D12S391, D2S441) included in the new European Standard Set (ESS) were calculated in a sample of 209 unrelated Italians with the Powerplex ESI® 17 system (Promega Corporation, Madison, WI). Forensic and population indices were estimated.

Samples were collected from unrelated individuals in 19 different Italian regions following informed consent.

DNA was extracted from saliva by Chelex method [1].

A prototype version of the PowerPlex ESI® 17 (Promega Corporation, Madison, WI) was used to amplify individuals' DNA according to manufacturer's recommendations. This multiplex contains 17 loci of which five are novel STR loci (D22S1045, D10S1248, D1S1656, D12S391, D2S441) included in the new European Standard Set (ESS) [2,3]. The other loci include the amelogenin, D3S1358, D19S433, D2S1338, D16S539, D18S51, TH01, vWA, D21S11, D8S1179, FGA and SE33.

PCR products were analysed by capillary electrophoresis in an ABI 3130xl Genetic analyzer (Applied Biosystem, Foster City, CA). Allele assignment was carried out by comparison with reference sequenced ladders (Promega Corporation, Madison, WI).

Arlequin software ver 3.0 [4] was used to calculate allele frequencies, population pairwise genetic distances (F_{ST}), expected heterozigosity (He), observed heterozigosity (Ho), and also to assess departures from Hardy–Weinberg equilibrium. Statistical parameters of forensic interest (Power of Discrimination, Power of Exclusion and Matching Probability) were calculated using PowerStats v1.2 (Promega Corporation, USA) software package [5].

The laboratory participates in the quality control initiatives of the GEDNAP (German DNA Profiling) group [6] and [7] (http://www.gednap.org).

The whole genotype data set, allele frequencies and forensic indices are available as an e-component.

Deviation from Hardy–Weinberg equilibrium has been detected only for D18S51, even after a Bonferroni correction. The combined power of exclusion (PE) and power of discrimination (PD) for the sixteen studied loci were 0.999999935 and 0.999999999, respectively. Based on heterozygosity and polymorphic information content (PIC), SE33 may be considered as the most informative loci. The exclusion of this locus slightly reduced the PE estimate (0.999999555). The PD value is similar for those calculated on a different Italian population set using the PowerPlex16 multiplex system (Promega Corporation, Madison, WI), the Identifiler® kit (Applied Biosystems) and for the markers included in the US Combined DNA Index System (CODIS) [8–10], while for the PE, the value obtained with the PowerPlex ESI® 17 System (Promega Corporation, Madison, WI) is higher. No pair of loci resulted in being in significant linkage disequilibrium after Bonferroni correction.

Allelic frequencies for the novel five STRs included in the ESS were compared to previously published population data [11–14]. Italian frequencies were significantly different to populations from Korea, China, Malay and India, while they were not significantly different to Spain and Italian data (for D2S441 and D1S1656, respectively). Fst genetic distances are reported (e-component).

This paper follows the guidelines for publication of population data requested by the journal [15].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2010.01.006.

References

[1] P.S. Walsh, D.A. Metzger, R. Higuchi, Chelex 100 as a medium for the simple extraction of DNA for PCR-based typing from forensic materials, Biotechniques 10 (1991) 506–513.
[2] P. Gill, L. Fereday, N. Morling, P.M. Schneider, The evolution of DNA databases—recommendations for new European STR loci, Forensic Sci. Int. 156 (2006) 242–244.
[3] M.D. Coble, J.M. Butler, Characterization of new miniSTR loci to aid analysis of degraded DNA, J. Forensic Sci. 50 (2005) 43–53.
[4] L. Excoffier, G. Laval, S. Schneider, Arlequin ver. 3.0: an integrated software package for population genetics data analysis, Evol. Bioinform. (Online) 1 (2005) 47–50.
[5] A. Tereba, Tools for analysis of population statistics, Profile in DNA (1999) 14–16 (free software distributed by the authors at website http://www.promega.com/geneticidtools/).
[6] S. Rand, M. Schurenkamp, B. Brinkmann, The GEDNAP (German DNA profiling group) blind trial concept, Int. J. Legal Med. 116 (2002) 199–206.
[7] S. Rand, M. Schurenkamp, C. Hohoff, B. Brinkmann, The GEDNAP blind trial concept. Part II. Trends and developments, Int. J. Legal Med. 118 (2004) 83–89.
[8] F. Brisighelli, C. Capelli, I. Boschi, P. Garagnani, M.V. Lareu, V.L. Pascali, A. Carracedo, Allele frequencies of fifteen STRs in a representative sample of the Italian population, Forensic Sci. Int. Genet. 3 (2009) 29–30.
[9] S. Presciuttini, N. Cerri, S. Turrina, B. Pennato, M. Alu', A. Asmundo, A. Barbaro, I. Boschi, L. Buscemi, L. Caenazzo, E. Carnevali, D. De Leo, C. Di Nunno, R. Domenici, M. Maniscalco, G. Peloso, S. Pelotti, A. Piccinini, D. Podini, U. Ricci, C. Robino, L. Saravo, A. Verzeletti, M. Venturi, A. Tagliabracci, Validation of a large Italian Database of 15 STR loci, Forensic Sci. Int. 156 (2006) 266–268.
[10] L. Garofano, G. Lago, C. Vecchio, M. Pizzamiglio, C. Zanon, A. Virgili, L. Albonici, V. Manzari, B. Budowle, Italian population data on the polymarker system and on the five short tandem repeat loci CSF1PO, TPOX, TH01, F13B, and vWA, J. Forensic Sci. 43 (1998) 837–840.
[11] P. Martín, O. García, C. Albarrán, P. García, I. Yurrebaso, A. Alonso, Allele frequencies of six miniSTR loci (D10S1248, D14S1434, D22S1045, D4S2364, D2S441 and D1S1677) in a Spanish population, Forensic Sci. Int. 169 (2007) 252–254.
[12] M.S. Han, Y.S. Kim, H.J. Jin, J.J. Kim, K.D. Kwak, J.E. Lee, J.M. Song, W. Kim, Forensic genetic analysis of nine miniSTR loci in the Korean population, Leg. Med. (Tokyo) 11 (2009) 209–212.
[13] R.Y.Y. Yong, L.S.H. Gan, M.D. Coble, E.P.H. Yap, Allele frequencies of six miniSTR loci of three ethnic populations in Singapore, Forensic Sci. Int. 166 (2007) 240–243.
[14] D. De Leo, S. Turrina, M. Marigo, N. Tiso, G.A. Danieli, Italian population data for D1S1656, D3S1358, D8S1132, D10S2325, VWA, FES/FPS, and F13A01, Forensic Sci. Int. 123 (2001) 71–73.
[15] A. Carracedo, J.M. Butler, L. Gusmão, W. Parson, L. Roewer, P.M. Schneider, Publication of population data for forensic purposes, Forensic Sci. Int. Genet., in press.

1872-4973/$ – see front matter
doi:10.1016/j.fsigen.2010.01.006

Andrea Berti[1,2]
Reparto Investigazioni Scientifiche di Roma-Sezione di Biologia- Viale di Tor di Quinto 119, Rome, Italy

Francesca Brisighelli[a,b,1,*]
[a]*Department of Zoology, University of Oxford, Oxford OX1 3PS, UK*
[b]*Institute of Legal Medicine, Genomics Medicine Group, University of Santiago de Compostela, Santiago de Compostela, Spain*

Alessandro Bosetti
Promega Italia S.r.l. Via Decembrio 28, 20137 Milano, Italy

Elena Pilli[3]
Department of Evolutionary Biology Laboratory of Anthropology Molecular Anthropology/Paleogenetics Unit University of Florence Via del Proconsolo 12 - 50122 Florence, Italy

Ciro Trapani[2]
Reparto Investigazioni Scientifiche di Roma-Sezione di Biologia- Viale di Tor di Quinto 119, Rome, Italy

Valentino Tullio[2]
Reparto Investigazioni Scientifiche di Roma-Sezione di Biologia- Viale di Tor di Quinto 119, Rome, Italy

Cristiano Franchi[2]
Reparto Investigazioni Scientifiche di Roma-Sezione di Biologia- Viale di Tor di Quinto 119, Rome, Italy

Giampietro Lago[2]
Reparto Investigazioni Scientifiche di Roma-Sezione di Biologia- Viale di Tor di Quinto 119, Rome, Italy

Cristian Capelli
Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

[*]Corresponding author at: Institute of Legal Medicine, University of Santiago de Compostela, rua san francisco s/n, Santiago de Compostela, Spain
E-mail addresses: fbrisi@libero.it, francesca.brisighelli@zoo.ox.ac.uk (F. Brisighelli)

[1]Both authors contributed equally to this work.
[2]Tel.: +39 0680980332; fax: +39 0680980336.
[3]Tel.: +39 055 2743032; fax: +39 055 2743038.

11 December 2009

3.3 Ancient DNA

Article 14. A nuclear DNA phylogeny of the woolly mammoth (*Mammuthus primigenius*). *Mol Phylogenet Evol.*

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 40 (2006) 620–627
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Short communication

A nuclear DNA phylogeny of the woolly mammoth (*Mammuthus primigenius*)

Cristian Capelli [a,1], Ross D.E. MacPhee [b,1], Alfred L. Roca [c,1], Francesca Brisighelli [a], Nicholas Georgiadis [e], Stephen J. O'Brien [d], Alex D. Greenwood [b,f,g,*]

[a] *Istituto di Medicina Legale, Universitá Cattolica del Sacro Cuore, Rome, Italy*
[b] *Division of Vertebrate Zoology, American Museum of Natural History, New York, New York, USA*
[c] *Laboratory of Genomic Diversity, Basic Research Program, SAIC-Frederick, Frederick, MD 21702, USA*
[d] *Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702, USA*
[e] *Mpala Research Center, P.O. Box 555, Nanyuki, Kenya*
[f] *GSF-National Research Centre for Environment and Health, Institute of Molecular Virology, Neuherberg, Germany*
[g] *Technical University of Munich, Institute of Virology, Munich, Germany*

Received 2 November 2005; revised 28 February 2006; accepted 7 March 2006
Available online 2 May 2006

1. Introduction

Although the woolly mammoth (*Mammuthus primigenius*) is one of the most intensively studied extinct species at the DNA level, mitochondrial DNA (mtDNA) markers have failed to unambiguously resolve its phylogenetic affiliation within Elephantidae. Most mtDNA-based elephantid phylogenies associate mammoths with African elephants (*Loxodonta africana* and *Loxodonta cyclotis*) to the exclusion of the Asian elephant (*Elephas maximus*) (e.g., Debruyne et al., 2003; Noro et al., 1998). However, other mtDNA studies (Ozawa et al., 1997), including recent sequencing efforts that yielded the complete mitochondrial genomes of two woolly mammoths (Krause et al., 2006; Rogaev et al., 2006), suggested that the Asian elephant is the closest living affine of mammoths. However, relationships inferred from mtDNA may be misleading due to the absence of a closely related outgroup species, or to the radiation of the three elephantid genera in rapid succession, which can produce discordance between a species tree and a gene (mtDNA) tree due to lineage sorting processes. Another difficulty is that in certain species—including elephants—the presence of nuclear insertions of mitochondrial sequences (Numts) can make identifying organellar mtDNA problematic (Greenwood and Pääbo, 1999; Thalmann et al., 2004). Moreover, Numt sequences are a routine, if unwanted, result of the procedures used in ancient DNA studies (Greenwood et al., 1999). Recently, cytonuclear genomic dissociation has been observed in African elephants, likely due to past hybridization between species (Roca et al., 2005). The existence of such dissociation phenomena could also confound mtDNA analysis within or among other elephantid species.

To date, the only extinct elephantid that has been amenable to confirmable molecular analysis by multiple research groups working with different specimens is the woolly mammoth (for a recent summary, see Greenwood, 2001). Yet, given the lack of consistent results across mtDNA phylogenetic studies, and given the possibility of discrepancies between the mtDNA tree and the species tree due to lineage sorting processes or to cytonuclear dissociation, nuclear DNA offers an alternative approach to studying woolly mammoth phylogeny. Nuclear DNA sequences from mammoths and other well-preserved extinct megafauna have been reported (Greenwood et al., 1999; Greenwood et al., 2001; Poinar et al., 2003; Poinar et al., 2006), and in principle it should be possible to characterize mammoth nuclear DNA sequences for the purpose of phylogenetic analysis. Of additional relevance, several nuclear genes have been investigated in a large number of individuals from different populations of *E. maximus*, *L. africana*, and *L. cyclotis* for the purpose of identifying fixed differences among groups and to establish their phylogenetic relationships (Roca et al., 2001). We have exploited and expanded this dataset to characterize the regions encompassing fixed differences among modern elephants in an

* Corresponding author. Fax: +49 (0) 89 31873329.
E-mail address: greenwood@gsf.de (A.D. Greenwood).
[1] These authors contributed equally to the study.

1055-7903/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.ympev.2006.03.015

effort to better ascertain the relationship of *M. primigenius* to extant elephantids.

2. Materials and methods

2.1. Samples

Two mammoth samples were included in this study. The first, from Engineer Creek, Alaska, has a radiocarbon date of 13,775 ± 145 years before present; nuclear and mitochondrial sequences for this specimen have been verified independently in different laboratories (radiocarbon dating described in Greenwood et al., 1999). Additional sequences have been reported for this mammoth (Binladen et al., 2006; Greenwood et al., 2001). The second sample, from Naskhok River in northeastern Wrangel Island (East Siberian Sea), has been dated to 4050 ± 40 years before present (Beta-195059; d13 corrected). For extant elephantids, our methods of sample collection, DNA extraction, PCR and sequencing have been previously described (Georgiadis et al., 1994; Roca et al., 2001, 2005).

2.2. Ancient DNA (aDNA)

Amplifications and re-amplifications were performed as described by Greenwood et al. (1999). To avoid contamination, processing was carried out in different research institutes: woolly mammoth samples were processed at the Istituto di Medicina Legale (Rome, Italy) while modern elephant samples were analyzed at the Laboratory of Genomic Diversity (Maryland, USA). Each aDNA amplification was performed in duplicate, cloned into a pGEM-T vector (Promega), transformed into electroporation competent bacteria, and five insert-positive clones per amplification sequenced to determine the consensus sequence of the clones (see Supplemental Figures 1–5 for all clone sequences used to derive the consensus sequences used in this study). PCR products ranged from 100 to 180 bp in length. Primer combinations used were, *CHRNA1*: L1 5′ GTTTAGTAGGTTGACTTCCA, R1 5′ GGACTCCATT ATGATCTTTA, L2 5′ GTGATGCACAGCATGAAC AT, R2 5′ AGCAGTTCGAATCCACCAGG, *GBA*: L 5′ GTAACCACTATGCTCCTCA, R 5′ CAGCCCTGAGG ACATCCAC, *BGN*: L1 5′ CTGAGCGCTAGGGCCAT CCA, R1 5′ ATGATGTTGCTGTGCAACA, L2 5′ TCAC ATCCACCAGTACAAAG, R2 5′ GTCTGTTTTAAAG CCTTTCC, *LEPR*: L 5′ TTATGGACTCTATATTGG AG, R 5′ TTGGTTGACCATCTGCAAGT. *VWF* sequences were taken from Greenwood et al. (1999).

2.3. Modern DNA

Genomic DNA (~50 ng) underwent amplification by PCR using 200 nM final concentration of each oligonucleotide primer in 1.5 mM $MgCl_2$, with AmpliTaq-GOLD DNA Polymerase (Applied Biosystems Inc. [ABI]). Primers were as previously reported for *BGN*, *CHRNA1*, and *GBA* (Lyons et al., 1997; Roca et al., 2001, 2005), but rock hyrax (*Procavia capensis*) *BGN* was amplified using new primers BGN-F1f (5′-AAGATCTCCAAGATCCAYGAGAARG) with BGN-R1f (5′-CCCARCCTGTACARCTTGGAGT A). *LEPR* used LEPR-F (5′-CCAAACCTCGAGGAAA GTTTACC) with LEPR-R (5′-AGGCTGCTCCTATGA TACCTCAA) for elephants and LEPR-F2 (5′-GCAGTG TACTGCTGCAATGA) with LEPR-R2 (5′-TGCAAAGT GCTTCCCACA) for hyrax. *VWF* was amplified using either vWF-F1a (5′-GATGGTGTCAACCTCACCTGT) or vWF-L1 (above) with vWF-R1a (5′-CAATGCCCACC GGGATCA); hyrax used vWF-F1a with vWF-R1a. For all primer pairs, PCR consisted of an initial 95 °C for 9:45 min; with cycles of 20 s at 94 °C, followed by 30 s at 60 °C (3 cycles); 58, 56, 54, or 52 °C (5 cycles each temperature); or 50 °C (last 22 cycles), followed by 75 s extension at 72 °C; with a final extension of 3 min at 72 °C. Sequences of several genes had been previously generated for multiple individuals of *E. maximus*, *L. africana*, and *L. cyclotis* (Roca et al., 2001, 2005), while novel elephant, mammoth and hyrax sequences generated for this study have been deposited in GenBank (*BGN*: DQ265804–DQ265820; *CHRNA1*: DQ265821–DQ265838; *GBA*: DQ265839–DQ265855; *LEPR*: DQ265856–DQ265888; *VWF*: DQ265889–DQ265919; Wrangel Island mammoth *BGN*: DQ267154, CHRNA1: DQ267155, DQ267156).

2.4. Phylogenetic analyses

Sequences were aligned using ClustalX (Thompson et al., 1997) and visually inspected. Two datasets were analyzed, each with concatenated DNA sequences from the genes *BGN*, *CHRNA1*, *GBA*, *LEPR* and *VWF*. The first dataset included sequences from elephantids and hyrax; 22 bp of the alignment in the 3′ fragment of *BGN* was excluded due to saturation of the region between hyrax and elephantids. The 3′ fragment of *CHRNA1* was an AfroSINE (Nikaido et al., 2003) present only in each of the elephantids; in hyrax it was coded as gaps and, to maximize resolution within elephantids, the maximum parsimony (MP) analysis treated gaps as a fifth state. The second dataset excluded the hyrax and used only elephantid sequences, including the complete 3′ sequence of *BGN*. In both datasets, a deletion (AAACC) was present in *CHRNA1* in one of the chromosomes (i.e., heterozygous) of elephant DS1534 and both chromosomes (homozygous) of LO3508; the deletion was part of the AfroSINE and removed from the alignment to avoid spurious affinity with hyrax. In a poly-T region of *LEPR* there was deletion of a thymine (in LO3505) or addition of a thymine (in LO3517); in each case the indel was present in only one of the chromosomes (heterozygous), and was not coded for analysis. These indels were present only in forest elephants and would not affect relationships inferred among elephantid genera. Modeltest 3.7 (Posada and Crandall, 1998) was used to determine the Akaike Information Criterion model of DNA sequence evolution that best fit the data; the model was implemented

for Neighbor Joining (NJ) and maximum likelihood (ML) methods in PAUP*4.0b10 (Swofford, 2002); MP was also run. Heuristic searches used 50 replicates of random taxon-addition and tree bisection-reconnection (TBR) branch swapping. Bootstrap resampling support was based on at least 100 replicates, with TBR branch swapping of starting trees obtained by stepwise addition. The model of evolution selected by Modeltest for each dataset was as follows. "Base" indicates the base frequencies for A, C, and G, with T inferred. "Nst" lists the number of substitution types listed in a rate matrix; the number of unique types may be inferred. "Rmat" is the rate matrix. "Rates" indicates the distribution of rates at variable sites. "Pinvar" indicates the proportion of invariant sites. For the elephantids + hyrax dataset: Lset Base = (0.2901 0.2364 0.2236) Nst = 6 Rmat = (1.0000 1.9849 0.3926 0.3926 3.7872) Rates = equal Pinvar = 0. For the elephantids-only dataset: Lset Base = equal Nst = 6 Rmat = (1.0000 1.3818 0.2787 0.2787 3.2657) Rates = equal Pinvar = 0. Tree scores are indicated on the Fig. 1 legend.

A Kishino Hasegawa (KH) test was run in PAUP* (Kishino and Hasegawa, 1989) using the following

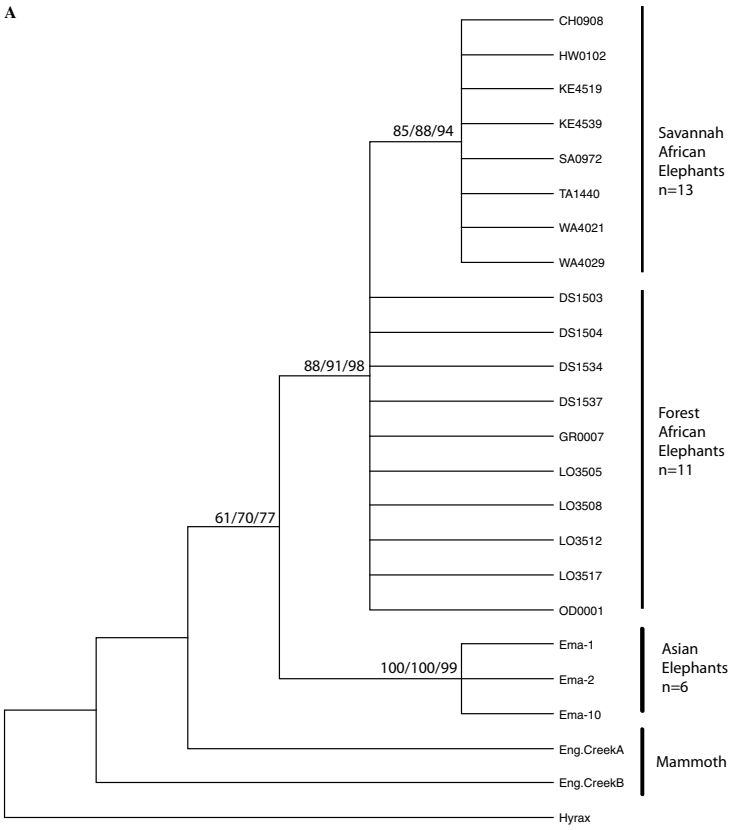


Fig. 1. Phylogenetic trees showing relationships among mammoths and living elephantids using DNA sequences from five nuclear genes (*BGN, CHRNA1, GBA, LEPR, VWF*). For both trees, bootstrap scores above 50% are shown for (left to right) maximum parsimony, Neighbor Joining, and maximum likelihood methods; "ns" indicates less than 50% bootstrap support for a given method. Modern elephant designations are taken from Table 1. Numbers indicated by species labels reflect the presence of additional individuals with duplicate sequences, not listed on the tree but shown in Table 1. (A) Strict consensus of 211,697 equally parsimonious trees produced by maximum parsimony analysis of 701 bp using hyrax as an outgroup, excluding a saturated portion of the 3′*BGN* sequence and treating gaps as a fifth state. The same interspecies relationships were suggested by MP (length 295; CI 0.990; RC 0.950), NJ (ME-score = 0.22282) and ML (-Ln likelihood = 1448.6734) methods. (B) The NJ tree, midpoint rooted for a 677 bp alignment excluding the hyrax sequence. The same interspecies relationships were suggested by MP (number of trees = 1000 [maxtrees], Length 62; CI 1.000; RC 1.000), NJ (ME-score = 0.03703) and ML (-Ln likelihood = 0.03758) methods.

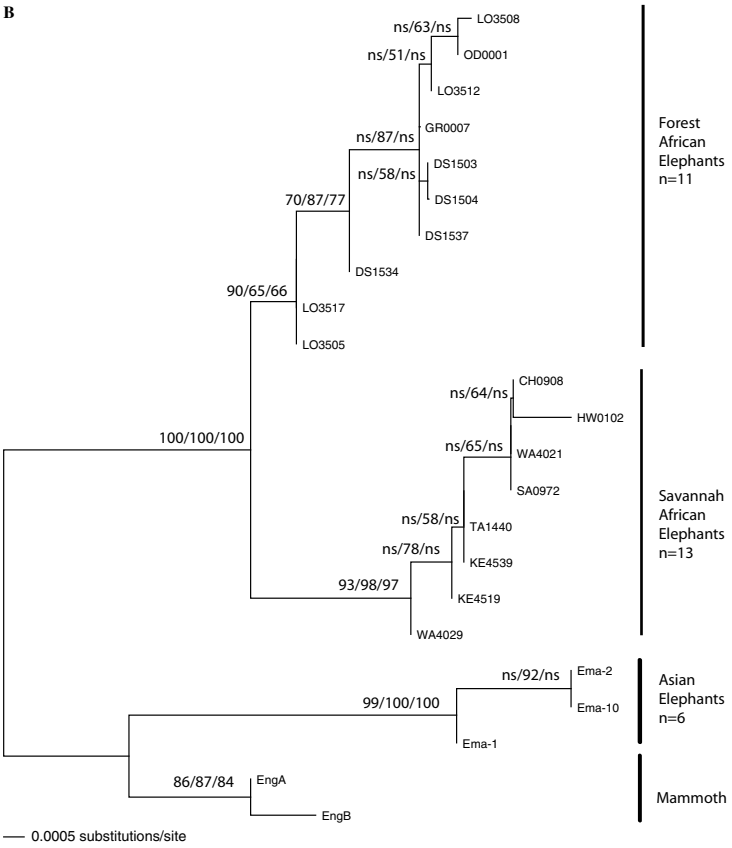


Fig. 1. (*continued*)

tree based on MP analysis of the dataset: (((((((CH0809, HW0102,SA0972,WA4021),KE4519,KE4539,TA1440, WA4029),(DS1503,DS1504,DS1537,GR0007,(LO3508, OD0001),LO3512),DS1534),LO3517),LO3505),(Ema-1, (Ema-2, Ema-10))),(Eng.CreekA,Eng.CreekB)),hyrax). This tree was compared to two other trees, both with the same intra-generic but different inter-generic relationships among the individuals. In the first tree used in both KH tests, *Loxodonta* and *Elephas* formed a clade excluding *Mammuthus*; in one comparison tree *Loxodonta* and *Mammuthus* were grouped to form a clade excluding *Elephas*, while in the other *Elephas* and *Mammuthus* formed a clade excluding *Loxodonta*.

Minimum spanning tree analysis was performed for the aligned elephantid sequences (without hyrax) using the TCS program (Clement et al., 2000).

3. Results

Two individual mammoths were genotyped at multiple nuclear DNA loci chosen for the potential presence of fixed nucleotide differences between *Elephas* and *Loxodonta*. A total of 681 bp of mammoth sequence was determined for loci *BGN* (175 bp), *CHRNA1* (193 bp), *GBA* (62 bp), *LEPR* (137 bp), and *VWF* (114 bp), with sequences for *BGN*, *CHRNA1*, and *VWF* amplified in two non-overlapping fragments. The mammoths were from different continents (Wrangel Island in northeastern Asia and Engineer Creek in Alaska) and chronologically separated by thousands of years. Thus, recovered sequences are likely to be minimally representative of geographic variation among mammoths for the loci characterized. In addition, little variation among mammoths has been

observed for cyt *b* (Debruyne et al., 2003). However, for nuclear loci the scale of variation would have to be determined by examining the sequences from additional mammoths.

The Wrangel Island mammoth specimen yielded lower quality DNA than the Alaskan sample and only produced replicable sequence for one *BGN* fragment and both *CHRNA1* fragments. It also yielded sequence in a single attempt to amplify *GBA*. For all fragments in common between the two mammoths, sequences were identical. For subsequent phylogenetic and network analyses, the Alaskan mammoth sequence was used.

In addition to the two mammoths, samples from six Asian elephants, 11 African forest elephants, 13 African savannah elephants, and one hyrax were characterized for all loci (Table 1 and Supplementary Table 1). Attempts to amplify the same genes in a manatee were unsuccessful. Both variable and fixed among-species elephantid differences were identified, with the mammoths exhibiting three unique polymorphisms, Asian elephants eight, African forest elephants ten, and African savannah elephants six (excluding indels). Transitions outnumbered transversions and indels were present in three of the genes including a homozygous 5 bp deletion in forest elephant LO3508 (Table 1). A single West African forest elephant, SL0001 from Sierra Leone, proved identical in combined gene sequences to LO3517 from Gabon in Central Africa (Table 1). This fails to confirm the suggestion of Eggert and colleagues (2002), based on mtDNA, that West African elephants comprise a separate species distinct from *L. africana* and *L. cyclotis*, although more specimens and nuclear markers would be required to confirm our result.

A 22 bp segment of the 3′ *BGN* elephantid alignment could not be aligned to hyrax sequence due to saturation and was removed. The remaining hyrax–elephantid alignment was 701 bp in length (21 parsimony informative sites), which included a 3′ *CHRNA1* fragment coded as gaps in the hyrax since it comprised part of an AfroSINE insertion present only in the elephantids (Nikaido et al., 2003). Phylogenetic analysis using hyrax as the outgroup suggested that *M. primigenius* is the most primitive elephantid with a subsequent branching of *Elephas* and the two *Loxodonta* species (Fig. 1A and Supplementary Figure 7). However, there are several

Table 1
Variable sites in nuclear genes among elephantids

		5′				*BGN*						3′							5′			*CHRNA1*		3′			*GBA*				*LEPR*									5′			*VWF*		3′	
		19	20	49	83	22	24	30	31	26	56	63	82	16	67	03	10	10	39	58	72	07	81	31	124	47	93	37
Mammoth	Eng.CreekA	C	C	G	T	C	G	G	A	G	G	T	G	G	T	T	T	AAACC	A	A	C	C	C	C	G	T	A	A	A	G	A	C	G	C	A	A	G
	Eng.CreekB	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
	Ema-1	.	.	.	.	.	T	.	.	.	.	.	C/G	A	C	.	C/T	.	G	.	T	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.
	Ema-2	.	.	.	.	.	T	.	.	.	.	.	C	A	C	.	C	.	G	.	T	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.
Asian	Ema-6	.	.	.	.	.	T	.	.	.	.	.	C/G	A	C	.	C/T	.	G	.	T	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.
Elephants	Ema-7	.	.	.	.	.	T	.	.	.	.	.	C/G	A	C	.	C/T	.	G	.	T	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.
	Ema-9	.	.	.	.	.	T	.	.	.	.	.	C	A	C	.	C	.	G	.	T	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.
	Ema-10	.	.	.	.	.	T	.	.	.	.	.	C	A	C	.	C	.	G	.	T	.	.	T	.	.	C	.	.	.	A/G	.	.	.	.	.	.	.
	DS1503	A	A	A	C	.	.	T	G	-	-	.	.	.	C	.	.	.	G	.	.	.	.	.	.	.	.	C	.	T	.	.	.	.	.	T	.
	DS1504	A	A	A	C	.	.	T	G	-	-	.	.	.	C	.	.	.	G	.	.	.	C/T	.	.	.	.	C	.	T	.	.	.	.	.	T	.
	DS1534	A/C	A/C	A	C	.	.	T	G	-	-	.	.	.	C	.	.	.	G	.	.	.	C/T	.	.	C/T	.	C	.	T	.	.	.	.	A/G	T	A/G
Forest	DS1537	A	A	A	C	.	.	T	G	-	-	.	.	.	C	.	.	.	G	.	.	.	.	.	.	C/T	.	A/C	.	T	.	.	.	.	.	T	.
African	LO3505	.	.	A	C	.	.	T	G	-	-	.	.	.	C	.	.	.	G	.	.	.	.	.	.	C/T	.	A/C	.	T	.	.	.	.	.	T	.
Elephants	LO3508	A	A	A	C	.	.	T	G	-	-	.	.	.	C	.	.	-----	G	.	.	.	.	.	.	C	.	C	.	T	.	.	.	.	.	T	.
	LO3512	A	A	A	C	.	.	T	G	-	-	.	.	.	C	.	.	.	G	.	.	.	C/T	.	.	C/T	.	C	.	T	.	.	.	.	.	T	.
	LO3517	.	.	A	C	.	.	T	G	-	-	.	.	.	C	.	.	.	G	.	.	.	C/T	.	.	C/T	.	A/C	.	T	.	.	.	.	.	T	.
	OD0001	A	A	A	C	.	.	T	G	-	-	.	.	.	C	C/T	.	.	G	.	.	.	.	.	.	C	.	A/C	A/G	T	.	.	.	.	.	T	.
	SL0001	.	.	A	C	.	.	T	G	-	-	.	.	.	C	.	.	.	G	.	.	.	C/T	.	.	C/T	.	A/C	.	T	.	.	.	.	.	T	.
	GR0007	A	A	A	C	.	.	T	G	-	-	.	.	.	C	.	.	.	G	.	.	.	C/T	.	A/G	C/T	.	A/C	.	T	.	.	.	.	A/G	T	A/G
	CH0908	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	A/G	.	T	.	.	.	.	.	C	.	T	.	.	.	T	.	T	.
	HW0102	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	G	.	T	.	.	.	.	.	C	.	T	.	.	.	T	.	T	.
	KE4519	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	T	.	.	.	.	.	C	.	T	.	.	.	.	.	T	.
	KE4539	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	T	.	.	.	.	.	C	.	T	.	.	.	C/T	.	T	.
Savannah	KR0014	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	T	.	.	.	.	.	C	.	T	.	.	.	C/T	.	T	.
African	KR0138	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	G	.	T	.	.	.	.	.	C	.	T	.	.	.	T	.	T	.
Elephants	NA4721	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	T	.	.	.	.	.	C	.	T	.	.	.	C/T	.	T	.
	SA0972	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	T	.	.	.	.	.	C	.	T	.	.	.	T	.	T	.
	SE2100	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	T	.	.	.	.	.	C	.	T	.	.	.	C/T	.	T	.
	TA1440	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	T	.	.	.	.	.	C	.	T	.	.	A/G	C/T	.	T	.
	WA4020	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	T	.	.	.	.	.	C	.	T	.	.	.	.	.	T	.
	WA4021	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	T	.	.	.	.	.	C	.	T	.	.	A/G	T	.	T	.
	WA4029	.	.	A	C	T	.	.	G	.	-	C	.	.	C	.	.	.	G	.	.	C/T	.	.	.	.	.	C	.	T	.	.	.	.	.	T	.
Hyrax	Pca-1	.	.	A	.	.	?	?	?	?	?	.	.	.	.	.	.	-	-	-	-	T	.	.	A	.	.	G	.	.	.	T	A	T	.	.	.

Eng.CreekA and B refer to the two alleles found in this specimen for VWF (Greenwood et al., 1999). IUPAC designations for bases are shown,"." represents identity to the reference sequence and "-" represent deletions or gaps. Positions saturated when aligned to the hyrax are indicated by a question mark "?". Mammoth character states not present in any elephants are shaded grey. The numbering begins from the first base after the 5′ primer for each mammoth sequence; for BGN, CHRNA1, and VWF, the 5′ and 3′ sequences are separated by a line and numbered separately. Modern elephant sample designations are taken from Roca et al., 2001. Localities for the *Loxodonta* samples are: DS-Dzanga Sangha in Central African Republic; GR-Garamba in Congo (Kinshasa); LO-Lope in Gabon; OD-Odzala in Congo (Brazzaville); SL-Sierra Leone; CH-Chobe and SA-Savuti in Botswana; HW-Hwange in Zimbabwe; KE-Central Kenya; KR-Kruger in South Africa; NA-Namibia; SE-Serengeti and TA-Tarangire in Tanzania; WA-Waza in Cameroon. Asian elephants were from zoos; those of known geographic origin derived from India (Ema-1), Sri Lanka (Ema-2 & 9) and Thailand (Ema-10).

reasons to be cautious with this conclusion. First, the removal of part of *BGN* from the analysis removes several sites variable among the elephantids. Second, the branch length leading from the outgoup is extremely long; in the analysis by Springer et al. (2005), which notably does not support the traditional definition of Tethytheria, the base of the paeungulate divergence (elephants (sirenians, hyraxes)) is dated to approximately 63 Ma. Very long outgroup branch lengths have been a difficulty for mtDNA based phylogenies of elephantids and are clearly a problem for nuclear DNA based analysis. As the more recent lineages having a common ancestry with extant elephants are all extinct, it remains to be seen if an appropriate outgroup species will allow for nuclear DNA study. A Kishino Hasegawa (KH) test compared the tree with the relationships suggested by MP analysis of the dataset, in which *Elephas* and *Loxodonta* form a clade excluding *Mammuthus* (Fig. 1A), to alternative trees in which intra-generic relationships were maintained but inter-generic relationships were altered. The KH test found that support for a tree with the *Elephas–Loxodonta* clade (Fig. 1A) was not significantly different from support for a tree with a *Loxodonta–Mammuthus* clade ($p = 0.32$) or support for a tree with an *Elephas–Mammuthus* clade ($p = 1.00$).

The analysis with hyrax had excluded a region of the *BGN* sequence that was saturated between elephantids and hyrax. To include the full elephantid sequence for *BGN*, a second analysis was run that excluded the hyrax and used all of the elephantid *BGN* sequence, along with the four other gene sequences, in an alignment 677 bp long (25 parsimony informative sites). The tree was mid-point rooted (Fig. 1B). The results demonstrated the expected separation of *L. cyclotis* and *L. africana* (Roca et al., 2001). Although the tree appears to suggest a slightly closer relationship of *Elephas* and *Mammuthus*, this interpretation should be treated with caution given the paucity of informative sites, the lack of an appropriate outgroup, and the possibility that some lineages may be accelerated. Nonetheless, a *Mammuthus–Loxodonta* association was not suggested by either analysis, in contrast to several mtDNA studies.

Network analysis could not distinguish between a closer association of woolly mammoths and Asian elephants or African elephants, with nine steps separating *Mammuthus* from the nearest *Elephas* individual, versus ten steps to the nearest *Loxondonta*. In particular, it is of interest to note (cf. Roca et al., 2001) that the distance between *L. cyclotis* and *L. africana* is high relative to the difference between either of these species and Asian elephants or woolly mammoths (4–13 steps, including indels, between *L. cyclotis* and *L. africana* versus 10–19 steps between *L. cyclotis* and *M. primigenius*, Fig. 2). The same analysis with gaps

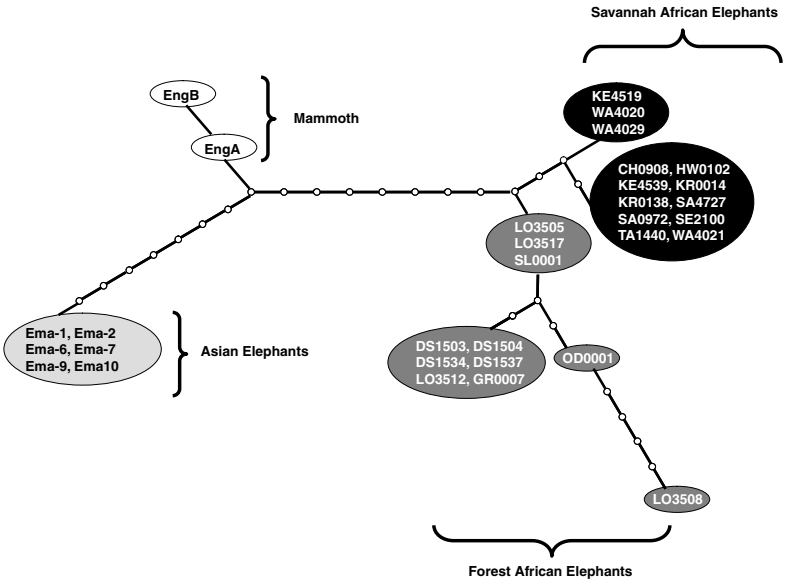


Fig. 2. Minimum spanning network of elephantid sequences. Open circles indicate the number of substitutions between nodes. Modern elephant designations are as in Table 1. The network depicted includes gaps as a fifth state. Heterozygous positions (Table 1) were scored as unknown. For example, position 107 in the *LEPR* gene was a C (cytosine) in all but six African elephants; these six individuals were heterozygous (C and A). Thus, some elephant individuals carried two different haplotypes with varying degrees of distance to sequences from other animals.

excluded yielded a similar network but reduced the number of steps along some branches (for example L03508 would be the same as OD0001; see also Table 1). The distance and diversity exhibited by *L. cyclotis* reflects a long history of reproductive isolation from *L. africana*.

4. Discussion

Although a small number of sites uniquely group woolly mammoths and Asian elephants, the phylogeny of the Elephantidae could not be resolved with the current dataset. However, the trend does not suggest a strong *Mamm uthus–Loxodonta* association as has been reported in several mtDNA based studies (Greenwood et al., 1999). By contrast, the *VWF* gene suggests a mammoth–*Elephas* association, as does the *BGN* gene. While *GBA* is ambiguous, *CHRNA1* favors a mammoth–*Loxodonta* association and *LEPR* slightly favors *Loxodonta–Mammuthus* as there are heterozygous forest elephant individuals with only one difference compared to mammoth while Asian elephants uniformly display two fixed *LEPR* differences versus the mammoth sequence. Nonetheless, none of our analyses combining all the sequences produced a *Mammuthus–Loxodonta* grouping.

The three elephantid genera radiated in quick succession in the late Miocene/early Pliocene (Maglio, 1973; Vignaud et al., 2002). Their evolutionary patterns may be comparable to that produced by the contemporaneous rapid radiation of the gorilla, chimpanzee and human lineages, in which the correct (gorilla (human, chimpanzee)) relationship is supported by only 60% of nuclear loci and phylogenetically informative sites, due to random sorting, recombination, genetic drift or homoplasy (O'hUigin et al., 2002; Satta et al., 2000). An added difficulty for interpreting elephantid relationships is that one target group is extinct. Lack of an appropriate outgroup sequence is another difficulty. Hyracoids and sirenians are the groups most closely related to proboscideans, but since their divergences occurred at the beginning of the Cenozoic 63 Ma, they are poor candidates for determining among-species branching patterns. Although mtDNA sequences have been reported for the mastodon (*Mammut americanum*), the results have not been independently replicated and nuclear DNA has never been retrieved from a mastodon (Yang et al., 1996).

Nonetheless, the results of this study suggest that further sequencing of woolly mammoth nuclear genes should resolve their phylogeny conclusively, although it will require a substantial increase in the number of informative sites and independent loci examined. Recent developments in sequencing technology suggest that whole genome analysis of extinct animals, particularly mammoths will be feasible (Poinar et al., 2006). We also conclude that the application of nuclear markers is now practicable and indeed preferable for systematic study of a wide variety of extinct animals represented by well-preserved remains in museum collections.

Acknowledgments

The authors thank Lars Giesen and Uwe Kohler (both of Medigenomix GmbH, Martinsried, Germany) for technical support. We are grateful to Claudia Englbrecht for critically reading the manuscript. We thank A. Brandt, S. Rosendale, and S. Mordensky for assistance. For elephant samples, we thank A. Turkalo, J.M. Fay, R. Weladji, W. Karesh, M. Lindeque, W. Versvelt, K. Hillman Smith, F. Smith, M. Tchamba, S. Gartlan, P. Aarhaug, A.M. Austmyr, Bakari, Jibrila, J. Pelleteret, L. White, M. Habibou, M.W. Beskreo, D. Pierre, C. Tutin, M. Fernandez, R. Barnes, B. Powell, G. Doungoubé, M. Storey, M. Phillips, B. Mwasaga, A. Mackanga-Missandzou, M. Keele, D. Olson, B. York, and A. Baker at the Burnet Park Zoo, M. Bush at the National Zoological Park, and A. Lécu at Zoo de Vincennes (Paris Zoo). We thank the governments of Botswana, Cameroon, the Central African Republic, Congo (Brazzaville), Congo (Kinshasa), Gabon, Kenya, Namibia, South Africa, Tanzania, and Zimbabwe for permission to collect samples. Samples were obtained in full compliance with specific Federal Fish and Wildlife Permits (endangered/threatened species and CITES Permits US 750138 and US 756611 to N.G.). For funding, we thank R. Ruggiero and the US Fish and Wildlife Service African Elephant Conservation Fund; and the National Geographic Society and European Union (through the Wildlife Conservation Society). This publication has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The work was financially supported by grants from the National Science Foundation (OPP 0117400), the Niarchos Fund, and the Evelyn Stefansson Nef Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2006.03.015.

References

Binladen, J., Wiuf, C., Gilbert, M.T., Bunce, M., Barnett, R., Larson, G., Greenwood, A.D., Haile, J., Ho, S.Y., Hansen, A.J., Willerslev, E., 2006. Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. Genetics 172, 733–741.

Clement, M., Posada, D., Crandall, K.A., 2000. TCS: a computer program to estimate gene genealogies. Mol. Ecol. 9, 1657–1659.

Debruyne, R., Barriel, V., Tassy, P., 2003. Mitochondrial cytochrome b of the Lyakhov mammoth (Proboscidea, Mammalia): new data and phylogenetic analyses of Elephantidae. Mol. Phylogenet. Evol. 26 (3), 421–434.

Eggert, L.S., Rasner, C.A., Woodruff, D.S., 2002. The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers. Proc. R. Soc. Lond. B Biol. Sci. 269, 1993–2006.

Georgiadis, N., Bischof, L., et al., 1994. Structure and history of African elephant populations: I. Eastern and southern Africa. J. Hered. 85, 100–104.

Greenwood, A.D., 2001. Mammoth biology: biomolecules, phylogeny, Numts, nuclear DNA, and the biology of an extinct species. Anc. Biomol. 3, 255–266.

Greenwood, A.D., Pääbo, S., 1999. Nuclear insertion sequences of mitochondrial DNA predominate in hair but not in blood of elephants. Mol. Ecol. 8, 133–137.

Greenwood, A.D., Capelli, C., Possnert, G., Pääbo, S., 1999. Nuclear DNA sequences from late Pleistocene megafauna. Mol. Biol. Evol. 16, 1466–1473.

Greenwood, A.D., Lee, F., Capelli, C., Desalle, R., Tikhonov, A.N., Marx, P.A., MacPhee, R.D.E., 2001. Evolution of endogenous retrovirus-like elements of the woolly mammoth (*Mammuthus primigenius*) and its relatives. Mol. Biol. Evol. 18, 840–847.

Kishino, H., Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29, 170–179.

Krause, J., Dear, P.H., Pollack, J.L., Slatkin, M., Spriggs, H., Barnes, I., Lister, A.M., Ebersberger, I., Paabo, S., Hofreiter, M., 2006. Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. Nature 439, 724–727.

Lyons, L.A., Laughlin, T.F., Copeland, N.G., Jenkins, N.A., Womack, J.E., O'Brien, S.J., 1997. Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nat. Genet. 15, 47–56.

Maglio, V.J., 1973. Origin and evolution of the Elephantidae. Trans. Am. Phil. Soc. Philad., New Series 63, 1–149.

Nikaido, M., Nishihara, H., Hukumoto, Y., Okada, N., 2003. Ancient SINEs from African endemic mammals. Mol. Biol. Evol. 20 (4), 522–527.

Noro, M., Masuda, R., Dubrovo, I.A., Yoshida, M.C., Kato, M., 1998. Molecular phylogenetic inference of the woolly mammoth *Mammuthus primigenius*, based on complete sequences of mitochondrial cytochrome *b* and 12S ribosomal RNA genes. J. Mol. Evol. 46 (3), 314–326.

O'hUigin, C., Satta, Y., Takahata, N., Klein, J., 2002. Contribution of homoplasy and of ancestral polymorphism to the evolution of genes in anthropoid primates. Mol. Biol. Evol. 19, 1501–1513.

Ozawa, T., Hayashi, S., Mikhelson, V.M., 1997. Phylogenetic position of mammoth and Steller's sea cow within Tethytheria demonstrated by mitochondrial DNA sequences. J. Mol. Evol. 44 (4), 406–413.

Poinar, H., Kuch, M., McDonald, G., Martin, P., Pääbo, S., 2003. Nuclear gene sequences from a late Pleistocene sloth coprolite. Curr. Biol. 13, 1150–1152.

Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R.D., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., Rampp, M., Miller, W., Schuster, S.C., 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311, 392–394.

Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14 (9), 817–818.

Roca, A.L., Georgiadis, N., Pecon-Slattery, J., O'Brien, S.J., 2001. Genetic evidence for two species of elephant in Africa. Science 293, 1473–1477.

Roca, A.L., Georgiadis, N., O'Brien, S.J., 2005. Cytonuclear genomic dissociation in African elephant species. Nat. Genet. 37 (1), 96–100.

Rogaev, E.I., Moliaka, Y.K., Malyarchuk, B.A., Kondrashov, F.A., Derenko, M.V., Chumakov, I., Grigorenko, A.P., 2006. Complete mitochondrial genome and phylogeny of Pleistocene mammoth *Mammuthus primigenius*. PLoS Biol. 4 (3), e73.

Satta, Y., Klein, J., Takahata, N., 2000. DNA archives and our nearest relative: the trichotomy problem revisited. Mol. Phylogenet. Evol. 14, 259–275.

Springer, M., Murphy, W., Eizirik, E., O'Brien, S.J., 2005. In: Rose, K., Archibald, J.D. (Eds.), The Rise of Placental Mammals: Origins and Relationships of the Major Extant Clades. Johns Hopkins Univ. Press, Baltimore, pp. 37–49.

Swofford, D.L., 2002. PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.0b10. Sunderland, Massachusetts, Sinauer.

Thalmann, O., Hebler, J., Poinar, H.N., Pääbo, S., Vigilant, L., 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. Mol. Ecol. 13, 321–335.

Thompson, J.D., Gibson, T.J., et al., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

Vignaud, P., Duringer, P., Mackaye, H.T., Likius, A., Blondel, C., Boisserie, J.R., De Bonis, L., Eisenmann, V., Etienne, M.E., Geraads, D., Guy, F., Lehmann, T., Lihoreau, F., Lopez-Martinez, N., Mourer-Chauvire, C., Otero, O., Rage, J.C., Schuster, M., Viriot, L., Zazzo, A., Brunet, M., 2002. Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. Nature 418, 152–155.

Yang, H., Golenberg, E.M., Shoshani, J., 1996. Phylogenetic resolution within the Elephantidae using fossil DNA sequence from the American mastodon (*Mammut americanum*) as an outgroup. Proc. Natl. Acad. Sci. USA 93 (3), 1190–1194.

Discussion

4. Discussion

4.1 Genetic variability in Sub-Saharan Africa

4.1.1 Cameroon

Given the complex background, Cameroon provides a unique opportunity to study how biological, geographic, and cultural factors interact in determining variation within and among human populations. Our study of microsatellite variation provides some useful insights into the back-to-Africa migration hypothesized on the basis of the occurrence of the R1b1*(xR1b1b2) Y-chromosomal haplogroup at high frequencies in North Cameroon. The reduced haplogroup variation already noticeable in North Cameroon, resulting from the high R1b1*(xR1b1b2) prevalence, was found to be associated to a noteworthy reduction of microsatellite diversity. The results of the search for undeclared family relationships suggest that the observed level of Y-chromosome diversity is probably the result of the specific demographic population history rather than the effect of a sampling bias, especially in Podokwo and Uldeme populations. Some robust signatures of isolation in the above-mentioned populations also for Y-chromosomal polymorphisms, including an extreme reduction of Y chromosomal intra-population variation and increase in inter-population diversity, were detected. Grouping of populations on a geographic basis, proved that all the genetic systems examined point to a greater variance of intra-population parameters and to a larger inter-population diversity for the northern part of the country. These findings may be explained by the fact that the two areas have been shaped by distinct peopling processes, which have been more complex for North Cameroon. This is also mirrored by the linguistic diversity, since the southern populations under study belong to the Benue Congo sub-branch of the Niger Kordofanian phylum (Niger Congo branch), whereas those from northern Cameroon belong to two different linguistic phyla (Afro-Asiatic and the Niger-Kordofanian).

Using autosomal STRs, the results neither clearly support nor contradict the role of language. Indeed, autosomal STRs provided the less clear and easily interpretable results among the genetic systems considered in this study

The lack of robust signals of correlation between paternally inherited polymorphisms and language or geography contrasts with previous studies carried out on larger areas of sub-Saharan Africa. Such a result should be considered in the light of the extreme reduction

of Y-chromosome intra-populational diversity (and the correlated increase of diversity among populations) observed in northern Cameroon populations.

By exploring patterns of genetic, geographic, and linguistic variation, we detect a preferential correlation between genetics and geography for mtDNA. This finding could reflect a female matrimonial mobility that is less constrained by linguistic factors than in males.

The different behavior of the two unilinearly transmitted polymorphisms may be viewed in the light of a model previously proposed. It integrates demographic and genetic aspects and incorporates ethnographic knowledge, identifying differences between HGs and FPs concerning direction of gene flow, extent of polygyny and respect of patrilocality as key factors for determining their diverse genetic structure. The results of the present study suggest the existence of a further heterogeneity among FPs regarding the ratio of female and male matrimonial mobility.

4.1.2 Western and Southern Africa

Whereas the dissection of single Y-chromosomal clades or subclades has helped to define the relationships between specific populations/groups, as well as reconstruct the demographic impact of migratory and cultural events, a wider and exhaustive phylogeographic analysis may indicate areas of the African continent where the extant human Y chromosome diversity first originated. Haplogroups A and B are ideal candidates for this task, given their distribution in Africa and the fact that they represent the earliest lineages to branch off within the Y chromosome genealogy. Previous analysis of the Y chromosome variation pointed to an South African/East African (SA/EA) origin following the identification of haplogroup A3b and, to a lower extent, B types in populations from these areas. However, our results clearly indicate that A3b branched later within haplogroup A, making it uninformative on the origin of the early human Y lineages. Haplogroup A is divided into two branches: A1, represented by western and central African types, and A2-A3, containing SA and EA chromosomes, with a few from central Africa. Haplogroup A2 is mostly composed of southern Africa types; however, an early branch in A2 is found in central Africa. Within haplogroup A3, A3b1, the southern Africa clade, is a sister clade to A3b2, common in eastern Africa, whereas A3a is only found among EAs. In haplogroup B, B2a and

B2b are two sister clades, whereas B*(xB2) aggregates a number of chromosomes from central Africa that were ancestral for the set of SNPs we tested. B2a has a very wide distribution and is mainly present in Bantu-speaking populations. Within haplogroup B2b, B2b* contains samples from eastern, south-eastern, and central Africa, with P6-derived chromosomes from South Africa and P7 types mainly from HG populations from central, eastern, and southern Africa. These results seem to indicate that southern Africa was an early destination of ancient human migrations from other regions other than the original source, which fails to support the hypothesis presented in a recent large-scale study of autosomal loci. With respect to the roles of eastern and central Africa, the data set presented here, although tentatively pointing toward a wide-scale preservation of ancient lineages in central Africa, is still compatible with a primary role for eastern Africa, in agreement with hypotheses generated from both mtDNA analysis and the study of the earliest *Homo sapiens* fossil remains.

4.2 Genetic variability in Europe: insight the Italian Peninsula

4.2.1 Anthropological implications

- When the climate improved and Paleolithic populations from European refugia repopulated the continent, some of the novel (or differently preserved) mtDNA and Y chromosome haplotypes also spread, giving rise to new star-like haplogroups in the phylogeny, marking the expansion range from each refugium.

The overwhelming importance of the Franco-Cantabrian refugium for the repeopling of much of Western and Northern Europe at the beginning of the Holocene has been obtained by the age estimates and geographic distributions of mtDNA haplogroups H1, H3, V, and U5b1b. Y chromosome haplogroups R1b1b2-M269, I1-M253, and I2b1-M223 support the important role of the Franco-Cantabrian refuge zone, whereas other Y haplogroups (I2a1-M423 and R1a1-M17) reveal that the Balkan and Ukrainian refuge zones were also major genetic sources for the human recolonization of Europe. In addition to these refugia mentioned above, another glacial refugium in Europe was the Italian Peninsula. However, neither mtDNA nor Y chromosome studies have yet been able to identify haplogroups marking expansions from this area, thus suggesting a marginal role, if any, of this southern European area in the postglacial repeopling of Europe.

Haplogroup U5 is one of the most ancient mtDNA haplogroups found in Europe. It evolved mainly within Europe where it spread after being involved in the first settlement of the continent by modern humans. Its phylogeny is characterized by two branches—U5a and U5b—which are common in most European populations, with U5b further split into U5b1 and U5b2. In 2006, a third uncommon branch, named U5b3, harboring the control region motif 16169A-16192-16235-16270-16519-150 was detected only in Sardinia, an island that remained unconnected with the mainland even when the sea level was lowest during the LGM and that was probably the last of the large Mediterranean islands to be colonized by modern humans.

Haplogroup U5b3 is virtually absent in the Near East and North Africa and is rare in Europe where, with the exception of the frequency peak in Sardinians (3.8%), its frequency barely reaches 1% only in some Mediterranean populations. Out of the 55 U5b3 mtDNAs detected in Sardinians are characterized by the diagnostic control-region motif of sub-haplogroup U5b3a1a, whose coalescence time estimate is between 4,600-6,300 years ago. This would mean that U5b3a1a has arisen in situ in Sardinia after the arrival of an U5b3a1 founder mtDNA from somewhere else in Europe and that U5b3a1a affiliation is a marker of maternal Sardinian ancestry (entry time: upper limit is 9,200–7,200 years ago (the age of U5b3a1 node), whereas the lower limit is 4,600-6,300 years ago (the age of the U5b3a1a node)), when the sub-haplogroup began to expand in Sardinia. This preliminary observation suggests a stronger link between Sardinia and southern France than with other European regions, including continental Italy. Archaeological data from the period 5,000–10,000 years ago show that the Monte Arci region of western Sardinia (Oristano province) was one of the four Mediterranean sources (together with the small islands of Palmarola, Lipari, and Pantelleria) of obsidian, the "black gold" of the Neolithic. Moreover, it has been calculated that the obsidian employed in the Neolithic sites of the southern France was almost exclusively from a "single" Monte Arci subsource, suggesting not only a preferential link between French sites and Sardinia but also preferential transport mechanisms, different from those connecting Sardinia with other Mediterranean regions (Corsica and northern Italy) where this selection of specific subsources has not been detected.

U5b3 haplogroup divergence is virtually identical to that reported for H1 and H3, thus indicating a population expansion at about the same time. Haplogroups H1 and H3

diffused from the Franco-Cantabrian refuge zone when climatic conditions improved; therefore, it is possible that also the founder U5b3 sequence expanded from the same area and the three haplogroups were involved in the same demographic processes. However, there is also an alternative scenario: the expansion of U5b3 could have still occurred at the same time as H1 and H3 when climatic conditions in Europe changed, but from a distinct geographical source. With consideration to the modern range distribution of U5b3, the only other potential candidate for the latter scenario is the glacial refuge in the Italian Peninsula.

- We extensively sampled along the peninsula to specifically address the issue of the distribution of Y chromosome genetic variation in the light of agriculture diffusion models. When compared to other European samples, no outgroups were found among the Italian samples. The samples in fact are distributed within the genetic variation shown by other European and Mediterranean populations. However, a limited degree of separation among Italian groups emerged along the first principal component, with Northern Italian samples closer to western European populations and Southern samples closer to South East and South Central European groups, with few exceptions. The Italian samples here analyzed appear to be placed within the ES–NW European cline that a number of previous studies have already described and that has been considered as compatible with a demographic scenario of admixture between the Near East farmers and the long-term European Mesolithic inhabitants. We have identified clines for three haplogroups (R1*(xR1a1), J2, E3b1), two of which showed also diversity gradients. In the light of Near Eastern gene flow, admixture analysis revealed Anatolian introgression in most of the Italian samples. These results support the Demic Diffusion (DD) model. Despite the presence of Neolithic genes in the current male Italian population, the admixture values as estimated by ADMIX suggested a differential impact of the newcomers across the Italian samples. The estimated degree of introgression is in fact not consistent across all areas, with Southern samples experiencing higher Anatolian contribution than Northern samples. Beside geographically different Near Eastern contributions, population replacement was not complete across the peninsula. It follows that both Neolithic and Mesolithic genetic components can be found in current Italian male gene pool. It is

interesting to note that haplogroup R1*(xR1a1) does show a frequency cline, opposite to the ones shown by J2 and E3b1, but apparently no diversity gradient is associated.

The Mesolithic populations had low population density and possibly limited gene flow across groups. If we assume that Mesolithic population were characterised by high frequencies of haplogroup R1*(xR1a1), genetic variation within this haplogroup would be independent of geographic sampling and instead mainly shaped by local demographic history. It follows that R1*(xR1a1) diversity would not be expected to show clines related to latitude but instead would be randomly distributed across populations.

The later newcomers as represented by Neolithic farmers, would have expanded and admixed with these Mesolithic groups, and generated, as expected, frequency and diversity clines along the direction of dispersal as indeed shown by their most representative chromosome types, E3b1 and J2.

The current set of data also provides a first frame for testing the hypothesis of genetic continuity from Palaeolithic to Mesolithic in Italy through the last Ice age. This would point to the presence of an Italian Pleistocene refugium, postulated for Iberian, Italian and Balkan Peninsula for a number of species, but not proposed for humans.

- A meta-analysis of Y-chromosome and mtDNA sequence data has been carried out in order to investigate patterns of genetic variation along Italy. Molecular indices indicate that most of the Italian samples show diversity values that are comparable to other European populations. However, some differences exist, especially in the isolated population of Ladins. Regional differences are much more evident when examining haplogroup frequencies in both uniparental markers. The differences are again more remarkable for the two linguistic isolates, the Ladins and Grecani Salentini.

The data shows that demographic movements during the Neolithic have left clinal latitudinal pattern along continental Italy. The contribution of the Neolithic into the Italian Peninsula has been estimated on the Y-chromosome as 14.5%% and in the mtDNA as the 10.5%. The Y-chromosome data show larger differentiation between North, Centre, and South than the mtDNA variation. AIMs show that the variation in Italians fits well with other European populations; and although these AIMs were not designed to allow a fine-grained resolution of intercontinental variation, they still recognize the presence of a minimal input coming from sub-Sahara. This Sub-Saharan contribution is also detected in

the uniparental makers. Bayesian-based admixture analysis on mtDNA data lead to the preliminary conclusion that North Africa contributed a total of 27% of the variation into Italy, while the Middle East contributed about 28%. However, these admixture estimates warrant further confirmation provided that these figures could be inflated due to the limited molecular resolution provided by the mtDNA control region data available. Italy shows patterns of molecular variation that mirror those of other European countries, although some heterogeneity exists as inferred from different analysis and molecular markers. From North to South, Italy shows clinal patterns that were most likely modulated during Neolithic times.

- A broadest analysis to date the spatial distribution of Y chromosome haplogroup M269, that can be split by R-S127 into European and western Eurasian lineages, was performed.

We see no relationship between diversity and longitude for R-M269. This correlation is the central tenet to the hypothesis that R-M269 was spread with expanding Neolithic farmers.

Dating of Y chromosome lineages is notoriously controversial the major issue being that the choice of STR mutation rate can lead to age estimates that differ by a factor of three (i.e. the evolutionary versus observed (genealogical) mutation rates).

In seeking to find a suitable set of STRs with which to estimate the average coalescence time (T) of sub-haplogroup R-S127, we have shown that not all STRs are of equal use in this context. We concentrated on estimating the duration of linearity (D) using different sets of STRs. Our analyses suggest that the D of an STR is key to its ability to uncover deep ancestry. Coalescence estimates explicitly depend on the STRs that one uses.

- On the basis of combined information of mt-SNPs and HVS-I sequence data, we confirmed that ~10% (26 individuals out of 258) of our Tuscans actually belong to one of the typical Near Eastern haplogroups, and have also a match with Near East populations. All of these Near East haplotypes are diverse (with the exception of those belonging to U7) and fall at the tips of the phylogeny, suggesting a recent arrival to the region. The typical Near Eastern U7 haplogroup occurs at relatively high frequency in the Elba Island

(~17%; 9 mtDNAs out of 53), and all of these U7 mtDNAs share the same HVS-I motif (T16271C-A16318T-T16519C), indicating that this lineage could represent a Near Eastern founder in the Isle. Complete genome sequencing of these nine U7 mtDNAs allowed the identification of a new subclade, U7a2a, characterized by transitions A13395G and T16271C. U7a2a is a sub-branch of U7a2. The amount of variation accumulated within U7a2a Etruscan cluster (assuming a single founder) can be dated in the range 1.1±0.1 to 2.3±0.4 kya BP, consistent with a recent arrival of this haplogroup to this Isle and compatible with the Etrurian culture (9th–1st century B.C.).

- To address the degree of historical NW African contribution, we used a combined Y-SNP-STR approach. The coalescent times for the three NW African specific haplogroups ranges between 5,000 and 24,000 years, spanning a number of historical scenarios each potentially explaining their presence on the Northern Mediterranean shores. It follows that estimating Medival North African (MNA) genetic legacy on the basis of haplogroups' occurrence only would be misleading. To avoid this limitation, we have extended our analysis to include STR data whose high mutation rate allows one to focus on more recent events. We screened more than 2300 South European samples to identify those haplotypes which are evolutionary close to NW African chromosomes. Total frequencies for these chromosomes range between 0 and 19% across southern Europe, the highest being in Cantabria and comprising a sample from the Pas Valley, previously shown to have an extremely high frequency of the North African haplogroup E1b1b1b.

Our estimates of NW African chromosome frequencies were highest in Iberia and Sicily, in accordance with the long-term Arab rule in these two areas. The chromosome frequencies in the two samples were not significantly different from each other (Fisher's exact test P= 0.83) but were both significantly different from the peninsular Italy sample (P=0.01). MNA types in the Italian peninsula, with at least a twofold increase over the Italian average estimate in three geographically close samples across the southern Apennine mountains (East Campania, Northwest Apulia, Lucera). When pooled together, these three Italian samples displayed a local frequency of 4.7%, significantly different from the North and the rest of South Italy (P=0.01), but not from Iberia and Sicily (P=0.12 and P=0.33, respectively). Arab presence is historically recorded in these areas following Frederick II's relocation of Sicilian Arabs. In Iberia, a non-random distribution might also

potentially be present, as suggested by our lower estimates in the northeast (Basque region and Catalans), but more samples across the peninsula will be required to properly address this issue. Assuming that a large population in regions such as Iberia, Sicily and Italy was present in the past, the *ratio* between Y chromosomes with a MNA ancestry and other types will have stayed approximately constant across time. Smaller areas, however, would have been influenced by drift, in the Pas Valley for example.

4.2.2 Forensic implications

4.2.2.1 *Y chromosome*

- A total of 186 haplotypes were defined by 9 Y-STRs haplotype comparisons. One hundred and thirty-five haplotypes of our collection (72.5% of the total) were not present in the previous database. So a large number of new types suggests that the Y chromosome diversity defined by our set of markers in Italy is only partially known and that a considerable sample-to-sample heterogeneity exists. Additional and extensive sampling would then be needed to obtain a more comprehensive description of the haplotype composition in the Peninsula.

Comparisons were conducted also using available published data on neighbour European and North African populations. AMOVA analysis revealed significant genetic variation across the three Italian areas (percentage of genetic variation across populations: 1.25, $p < 0.05$). No significant population differentiation was found by using Fisher's Exact Test. Significant variation was found when the three samples were grouped and tested against the GeFI database (AMOVA, percentage of genetic variation across populations: 0.24, $p < 0.05$).

- The goal of this Y chromosome study was to contribute to an Italian database of forensic interest given that: (i) most of the Y chromosome data available in literature is either restricted in the number of loci (minimal haplotype), (ii) generally focus on single populations, or (iii) cover the Italian peninsula only partially. Our sampling comprises individuals from North, Central and South of Italy and has been genotyped for the full set of Y-STRs considered in the AmpFlSTR Yfiler Amplification Kit (AB). We have shown that the analysis of full Y-STR profiles (Yfiler) vs minimal haplotypes is important for the

detection of local genetic differences, and that the genetic discrimination provided for the Yfiler profiles is substantiality larger than for the minimal haplotype.

We have performed AMOVA analysis in different population contexts, and we have demonstrated that measure of population stratification strongly depends not only on the amount of genetic information used, but much more importantly, on the amount of population samples employed in the analysis. This guarantees more caution when using FST values to correct for population stratification in forensic casework.

- DYS19 locus PCR doublets exist worldwide in humans. Asian individuals bearing this locus duplication belong to C3c haplogroup. In the course of a large Italian Y chromosome screening, we have identified 10 additional DYS19 duplication-bearing chromosomes. By additional typing work, we can report that all these chromosomes classify in the haplogroup G2*(xG2a,b). Our finding clearly implies that duplication at DYS19 had to occur at least twice, driven by as many independent genetic events. In the light of the current existence of the two lineages, the single-origin hypothesis implies that the original duplication event had to occur very early in the timescale of human evolution.

However, the currently available world-wide set of genotyped samples from different Y chromosome genealogy branches shows a limited number of DYS19 duplication bearing chromosomes, suggesting that, in the case of single origin, multiple reversion events should have occurred. The multiple-origin hypothesis, requesting so far a minimum of only two independent duplication events, appears, as a more parsimonious alternative, the most likely.

4.2.2.2 *Autosomal data*

- Fifteen autosomal STRs from PowerPlex16 multiplex system (Promega Corporation, Madison, WI) were analyzed in more than 400 unrelated samples from 9 different areas of Italy.

Population differentiation test showed no significant differences between populations. The AMOVA test was performed considering each population independently or grouped according to their geographic origin (North, Central and South of Italy). Results showed that most of the genetic variation occurs within individuals, underlining no evidence of

population structure. The combined power of exclusion (PE) and power of discrimination (PD) for the fifteen studied loci were 0.964708775 and 0.9999999997, respectively. The combined matching probability value was 1 in 3.33x108. Based on heterozygosity and polymorphic information content (PIC), FGA may be considered as the most informative loci. Locus-by-locus allelic frequencies were compared to previously published Italian and Mediterranean population data. After applying the Bonferroni correction for multiple tests, population differentiation tests showed that Italy had significant differences with Bosnia–Herzegovina in 4 out of 13 loci (D5S818, D7S820, D16S539 and CSF1PO), with Spain in 3 out of 13 loci (D3S1358, TH01 and D16S539) and with Kosovo Albanians in 1 out of 13 loci (D18S51).

- Allele frequencies of new STRs loci included in the PowerPlex ESI® 17 System (Promega Corporation, Madison,WI) were calculated in a sample of 209 unrelated Italians.

The combined power of exclusion (PE) and power of discrimination (PD) for the sixteen studied loci were 0.999999935 and 0.999999999, respectively. Based on heterozygosity and polymorphic information content (PIC), SE33 may be considered as the most informative loci. The exclusion of this locus slightly reduced the PE estimate (0.999999555). The PD value is similar for those calculated on a different Italian population set using the PowerPlex16 multiplex system (Promega Corporation, Madison, WI), the Identifiler1 kit (Applied Biosystems) and for the markers included in the US Combined DNA Index System (CODIS), while for the PE, the value obtained with the PowerPlex ESI® 17 System (Promega Corporation, Madison,WI) is higher. No pair of loci resulted to be in significant linkage disequilibrium after Bonferroni correction.

4.3 Ancient DNA. Mammoth phylogenetic affiliation.

- Although a small number of sites uniquely group woolly mammoths and Asian elephants, the phylogeny of the Elephantidae could not be resolved with the current dataset.

However, the trend does not suggest a strong *Mammuthus–Loxodonta* association as has been reported in several mtDNA based studies. By contrast, the VWF gene suggests a *Mammuthus–Elephas* association, as does the BGN gene. While GBA is ambiguous, CHRNA1 favors a *Mammuthus–Loxodonta* association and LEPR slightly favors *Loxodonta–Mammuthus* as there are heterozygous forest elephant individuals with only one difference compared to mammoth while Asian elephants uniformly display two fixed LEPR differences versus the mammoth sequence. Nonetheless, none of our analyses combining all the sequences produced a *Mammuthus–Loxodonta* grouping.

Conclusions

5. Conclusions

5.1 Sub-Saharan Africa

5.1.1 Cameroon

1) The combined effect of the propensity to inter-populational admixture of females, favored by cultural contacts (especially linguistic factors), and of genetic drift acting on Y-chromosomal diversity could account for the peculiar genetic pattern observed in northern Cameroon. Our results are in line with the hypothesis of a tendency towards isolation among *Montagnards* of northern Cameroon, related to physical barriers created by the mountainous environment.

5.1.2 Western and Southern Africa

2) Haplogroup A and B distribution is almost exclusively restricted to sub-Saharan Africa, with a peak of 65% in group of foragers. Specific lineages related to regional preagricultural dynamics in different areas of sub-Saharan Africa, contribute to understand the complex evolutionary relationship among African hunter-gatherers (Pygmies and Koisan speaking groups from Souther Africa). The role of southern Africa can now be seen more as sink than a source of the first migrations of modern humans from eastern and central part of the continent.

5.2 Genetic variability in Europe: insight the Italian Peninsula.

5.2.1 Anthropological conclusions

3) Holocene expansion (~11 kya) of haplogroup U5b3 occurred along the Mediterranean coasts, mainly toward nearby Provence (southern France). From there, ~7,000–9,000 years ago, a subclade moved to Sardinia (obsidian trade), leaving a distinctive signature in the modern people of the island. This scenario strikingly matches the age, distribution, and postulated geographic source of a

Sardinian Y chromosome haplogroup (I2a2-M26), a paradigmatic case in the European context of a founder event marking both female and male lineages.

4) The role of Neolithic farmers was greater than Greek historical colonizers of South Italy and this results support a male demic diffusion model (diversity distributed along a North–South axis), even if population replacement was not complete and the degree of Neolithic admixture with Mesolithic inhabitants was different in different areas of Italy; drift definitively had a role in shaping current Y chromosome genetic variation.

5) Italy shows clines of variation attributable to the demographic movements of first Paleolithic settlements, posteriorly modeled by the Mesolithic and to a lower extent Neolithic farmers. Regional differences arose with the time, which are more notable in linguistic isolates, such as the Ladins, and to a minor extent, the Grecani Salentini.

6) The opinion on the origin of the major European Y chromosome haplogroup R-M269, has moved away from Paleolithic origins to the notion of a younger Neolithic spread of this chromosome from the Near East. No geographical trends in diversity, in contradiction to expectation under Neolithic hypothesis, were observed, and suggest an alternative explanation for the apparent cline in diversity recently described. The frequency of R-M269 across Europe could be related to the growth of multiple, geographically specific sub-lineages that differ in different parts of Europe.

7) The investigation of a large and representative sample set and the analysis of complete mtDNA genomes support the hypothesis that Tuscany still preserves the fingerprint of a historical connection with the Near East. However, it should be stressed that this represents just a minor component of the Tuscan genetic make-up and suggests that historically different layers were superimposed over the Mesolithic gene pool of the Peninsula.

8) More than 56% of the Italian individuals identified here as having a recent NW African influence do not have a match in a large Italian Y chromosome dataset comprising almost 1200 individuals. Of these, 31% instead perfectly overlap with types from NW African populations, potentially providing misleading advice to investigators. Our results clearly confirm that conclusions based on single chromosomes should be taken very cautiously.

5.2.2 Forensic conclusions

9) A large number of new types suggests that the Y chromosome diversity defined by our set of markers (nine Y-STRs) in Italy is only partially known and that a considerable sample-to-sample heterogeneity exists. Additional and extensive sampling would then be needed to obtain a more comprehensive description of the haplotype composition in the Peninsula.

10) Although Italy seems to show lower levels of among population variance than other European regions, these values are high enough (about 3-11%) to guarantee caution when using the Y-chromosome test in forensics. Thus, the results raise the issue of developing local reference databases for forensic purposes given the substantial differences existing across the Italian Peninsula, in order to better understand the substructure of the country.

11) Y chromosome haplogroup and microsatellite-based network analysis, seem to indicate that DYS19 duplications do not have a commun origin but occurred more than once during human evolution by non-allelic homologous recombination or gene conversion.

12) A number of Italian autosomal databases are currently available. However, those are either restricted in the number of loci, focus on single populations or cover the Italian Peninsula only partially. Our sample, by comprising individuals from North, Central and South of Italy, and analyzing 15 loci (Powerplex 16 multiplex

system; Promega Corporation, Madison, WI) provides additional information on the genetic variation of the Italian population.

13) Forensic indices for the five novel autosomal STRs included in the Powerplex ESI® 17 system (Promega Corporation, Madison, WI) were calculated. The comparisons of the obtained results with other commercial kits (PowerPlex16 multiplex system, Identifiler1 kit, US Combined DNA Index System (CODIS), indicate that, depending on the STRs used, the values for the different indices used in paternity tests or criminal cases can change.

5.3 Mammoth and its phylogenetic affiliation

14) The results of the study suggest that further sequencing of woolly mammoth nuclear genes should resolve their phylogeny conclusively. Recent developments in sequencing technology suggest that whole genome analysis of extinct animals, particularly mammoths, will be feasible. We also conclude that the application of nuclear markers is now practicable and indeed preferable for systematic study of a wide variety of extinct animals represented by well-preserved remains in museum collections.

Resumen

6. **Resumen**

La genética de poblaciones humanas tiene por objeto entender las fuerzas que determinan la evolución, tratando de reconstruir la historia de las poblaciones humanas. La caracterización de la distribución de la variabilidad genética dentro de poblaciones diferentes, en las distintas regiones del mundo, permite investigar las afinidades genéticas, o incluso la proximidad filogenética, dentro y entre las proprias poblaciones. El estudio de como se distribuye la variabilidad genética entre y dentro de las poblaciones es un aspecto clave para los estudios de asociación en estudios de enfermedades, así como para la genética forense. En este último caso, el estudio de la variabilidad genética es esencial para la identificación individual a través de una "huella" genética, es decir, un conjunto de marcadores que puede ser tan variable que la combinación alélica observada se puede considerar casi individuo-específica.

Los principales objetivos de la genética de poblaciónes son: 1) describir la distribución de la diversidad genética de las poblaciones modernas y su distribución entre las sub-poblaciones, es decir, la estructura poblacional, 2) deducir los eventos prehistóricos e históricos que determinaron la diversidad observada y la estructura moderna.

También la genética forense, la ciencia que combina la genética de poblaciones y la medicina forense, está utilizando la variabilidad genética de los seres humanos. En este caso tenemos dos aplicaciones principales: 1) la identificación individual en los casos penales y 2) la identificación de dos personas como parientes cercanos. La composición genética de un individuo no puede considerarse en forma aislada, sino que tiene que estar relacionada con el grado y la estructura de la variación genética presente en la población a la que el individuo pertenece. Los métodos empleados de genotipado de ADN, de hecho, no pueden garantizar que el genotipo sea único y que no haya otra persona que lleve los mismos marcadores. Debido a esto, se calculan las probabilidades bajo determinadas hipótesis. La estimación de las probabilidades se basa en el conocimiento de las frecuencias de los genotipos de la población a la cual las personas involucradas en cada caso pertenecen.

La mayor parte del genoma se hereda de forma biparental y recombina. Sin embargo, dos segmentos particulares del ADN se heredan solamente de uno de los padres y no recombinan: el ADN mitocondrial y, para la mayor parte de su longitud, el cromosoma Y.

Sub-Saharan Africa

La difusión de las primeras comunidades de *Homo sapiens* moderno en el continente africano, después de su supuesto origen en el África oriental 150.000-200.000 años atrás, es un tema complejo que aún debe ser delucidado. La limitada disponibilidad de registros fósiles de *Homo sapiens* moderno en este continente es uno de los factores más limitantes para el estudio de este tema. Por otro lado, los procesos de poblaciónes más recientes que dieron forma al poblamiento de África sub-sahariana, se han estudiado, en los ultimos años, con mayor detalle. Esta area fue influenciada intensamente por la reciente expansión de las lenguas bantúes, y actualmente está habitada por unos de los últimos cazadores-recolectores de las comunidades Áfricanas, los pigmeos. África es una región importante para estudiar la diversidad genética humana debido a su compleja historia y a la dramática variación en el clima, la dieta y la exposición a enfermedades infecciosas, que se traducen en altos niveles de variación genética y fenotípica de las poblaciones africanas. Una mejor comprensión de los niveles y patrones de variación en los genomas africanos, junto con los datos sobre los rasgos de fenotipos variables, es fundamental para la reconstrucción de los orígenes del hombre moderno y de la base genética de la adaptación a diversos ambientes (Campbell and Tishkoff 2008).

África es una región de considerable diversidad genética, lingüística, cultural y fenotípica. Hay más de 2.000 diferentes grupos etno-lingüísticos en África, que hablan lenguas que constituyen casi un tercio de las lenguas del mundo (http://www.ethnologue.com/)(Campbell y Tishkoff 2008).

Europa: la península italiana

Debido a sus características geo-morfológicas, Italia fue uno destino de elección por los grupos humanos procedentes de África, Oriente Medio y otros lugares de Europa.

La presencia de homínidos en la península italiana ha sido compleja y prolongada en el tiempo. El *Homo sapiens* probablemente hizo su primera aparición en esta zona alrededor de hace 30.000-40.000 años (Cunliffe 2001). Alrededor de hace 11.000 años, en el Creciente Fértil nuevos recursos llegaron a estar disponible para los seres humanos en los medios de cultivos y animales domesticados. La nueva tecnología estaba ahora capaz de apoyar a las comunidades grandes y proporcionó los recursos para una expansión demográfica (Cunliffe 2001). La tecnología se extendió muy rápido a través de la península europea, llegando a los

límites occidentales sólo 4.000 años más tarde (Ammerman and Cavalli-Sforza 1984). El impacto demográfico relacionado es todavía un tema de debate, pero el consenso parece haber sido alcanzado en la contribución sustancial del Neolítico en la zona del Mediterráneo (Chikhi et al. 2002; Semino et al. 2000; Simoni et al. 2000).

ADN antiguo

Los primeros estudios de ADN antiguo han utilizado la clonación bacteriana para amplificar pequeñas secuencias recuperadas a partir de pieles de animales y momias humanas, y han revelado la ineficiencia de la cinética de reacción de esta técnica (Higuchi et al 1984;. Pääbo 1985, 1989). Estos estudios demostraron que el material genético presente en ejemplares antiguos es principalmente microbiano y que el ADN endógeno se limitaba generalmente a muy bajas concentraciones de fragmentos cortos y dañados de loci en múltiples copias, como el ADN mitocondrial. El enorme poder de amplificación de la PCR también ha creado una mayor sensibilidad a la contaminación con ADN moderno, y al mismo tiempo, una importante potencial fuente de contaminación a través de las concentraciones extraordinarias de los productos de PCR amplificados con anterioridad. Como consecuencia, los falsos positivos resultantes de contaminaciónes intralaboratorio siguen siendo un problema importante en la investigación de ADN antiguo. Una serie de estudios a gran escala han comenzado a revelar el verdadero potencial del ADN antiguo para registrar los métodos y los procesos de evolución, proporcionando una manera única de probar los modelos y las hipótesis utilizadas para reconstruir los patrones de la evolución, la genética de poblaciones y el cambio paleoecológico (Willerslev and Cooper 2005).

Objetivos

Los principales objetivos de esta tesis son los siguientes:

1. describir la distribución de la diversidad genética para las poblaciones humanas modernas y la distribución entre las subpoblaciones, con énfasis en las poblaciones de África subsahariana y Europa, y con un enfoque especial en Camerún, África occidental y central, e Italia, respectivamente. Hicimos uso de la información recopilada en los estudios previos, así como datos históricos, lingüísticos y geográficos.

2. inferir los eventos prehistóricos e históricos que determinaron la diversidad observada y la estructura en poblaciones actuales.

En la consecución de estos objetivos globales, esta tesis se divide en tres grandes grupos: dos grupos geográficos, incluyendo África y Europa, y un grupo de ADN antiguo. Los objetivos intermedios estan asì definidos:

i) **<u>África subsahariana</u>**

1) evaluar la variación genética de los microsatélites autosómicos y del cromosoma Y en una amplia serie de muestras de Camerún, teniendo también en cuenta factores geográficos y culturales.

2) caracterizar la variación genética del cromosoma Y con el fin de reconstruir los hechos demográficos e identificar linajes específicos relacionados con la difusión de las lenguas, la agricultura y el pastoreo en África subsahariana.

ii) **<u>Europa</u>**

1) realizar un análisis filogeografíca de la variación del ADNmt en el nivel más alto de resolución molecular, tratando de explicar cómo el área de refugio ubicada en la península italiana contribuyó a la recolonización humana del continente a principios del Holoceno.

2) investigar el papel de la península italiana como parte de un proceso más global del poblamiento de Europa, teniendo en cuenta las consecuencias demográficas de la revolución de la agricultura en la zona, mediante el genotipado de marcadores del cromosoma Y en un gran número de muestras.

3) analizar los patrones genéticos de Italia desde una perspectiva global, utilizando 12 diferentes poblaciones a lo largo de la península italiana, dos de ellas aislados lingüísticos, a través del analisis de la región control del ADN mitocondrial y de una selección de SNPs de la región codificante, un panel de SNPs del cromosoma Y y, además, AIMs autosómicos.

4) evaluar el origen del haplogrupo de cromosoma Y, R1b1b2-M269, a través de la investigación de los patrones de frecuencias y de diversidad, en una grande colección de muestras.

5) investigar el orígen de los etruscos (una de las más enigmáticas civilizaciones no indoeuropeas) a través del análisis de los toscanos modernos utilizando SNPs y la secuenciación completa del genoma del ADNmt
6) determinar la reciente contribución masculina del noroeste de África en la Península Ibérica y en Sicilia.
7) contribuir a enriquecer las bases de datos de cromosoma Y con unos conjuntos de datos de alta resolución del cromosoma Y
8) evaluar la correlación genealógica entre los cromosomas Y que llevan duplicación microsatélites en el sistema DYS19.
9) contribuir en las frecuencias alelicas de marcadores autosómicos de las diferentes poblaciones italianas.

iii) **ADN antiguo**
1) estudiar el mamut lanudo (*Mammuthus primigenius*), utilizando marcadores de ADN mitocondrial con el fin de resolver su afiliación filogenética dentro los elefántidos.

Discussion

La variabilidad genética en África subsahariana

Camerún

Nuestro estudio sobre la variación microsatélite ofrece algunas ideas útiles sobre la hipótesis de la migración de regreso a África, en base a la aparición en altas frecuencias del haplogrupo de cromosoma Y R1b1*(xR1b1b2), en el norte de Camerún.

El nivel observado de diversidad de cromosoma Y es probablemente el resultado de la historia demográfica específica de la población, y no el efecto de una parcialidad del muestreo, especialmente en las poblaciones de Podokwo y Uldeme. Agrupaciónes de las poblaciónes sobre una base geográfica, demostran que todos los sistemas genéticos examinados apuntan a una mayor varianza de los parámetros dentro de la población y una mayor diversidad entre la población de la zona norte del país. Estos resultados pueden explicarse por el hecho de que las dos áreas han sido formadas por procesos de poblamiento diferentes, más complejos para el Norte de Camerún. Esto también se refleja en la diversidad lingüística, ya que las poblaciones del sur bajo estudio pertenecen a la sub-rama Benué-Congo del filo Níger-

Kordofanian (rama Niger-Congo), mientras que las del norte de Camerún, pertenecen a dos diferentes phyla lingüísticos (afro-asiática y Níger-Kordofanian).

Al explorar los patrones de variación genética, geográfica y lingüística, se detecta una correlación preferencial entre la genética y la geografía en el ADNmt. Este hallazgo podría reflejar una movilidad matrimonial de las mujeres que está menos limitada por factores lingüísticos que en los hombres.

El diferente comportamiento de los dos polimorfismos a transmisión unilinear puede ser visto a la luz de un modelo propuesto anteriormente; de hecho, integra los aspectos demográficos y genéticos, e incorpora el conocimiento etnográfico, identificando las diferencias entre CRs y agricultores sobre la dirección del flujo genico, la extensión de la poligamia y el respeto de la patrilocalidad como factores clave para la determinación de su diferente estructura genética. Los resultados del presente estudio sugieren la existencia de una heterogeneidad mayor entre los agricultores con respecto a la relación entre la movilidad matrimonial femenina y masculina.

África occidental y meridional

La disección de clados o subclados del cromosoma Y ayuda a definir las relaciones entre las poblaciones y grupos específicos, así como la reconstrucción del impacto demografíco de los eventos migratorios y culturales. Un análisis más amplio y exhaustivo filogeográfico puede indicar las áreas del continente africano donde la diversidad humana del cromosoma Y se originó. Los haplogrupos A y B son los candidatos ideales para esta tarea, teniendo en cuenta sus distribuciónes en África y el hecho de que representan los primeros linajes de la genealogía que se ramificaron en el cromosoma Y.

Análisis anteriores de la variación de cromosoma Y señalaron un origen sudafricano /África Oriental (SA/EA), tras la identificación de los haplogrupos A3b y, en menor medida, B en las poblaciones de estas áreas. Sin embargo, nuestros resultados indican claramente que A3b ramificó más tarde, dentro del haplogrupo A, por lo que es poco informativo sobre el origen de los linajes Y humanos primitivos. El haplogrupo A se divide en dos ramas: A1, representada por los tipos de África occidental y central, y A2-A3, que contienen cromosomas SA y EA, con algunos de África central. Haplogrupo A2 se compone sobre todo de tipos de sur de África, también si una rama de A2 se encuentra en el centro de África. Dentro del haplogrupo A3, A3b1, el clado de sur de África, es un clado hermano de A3b2,

común en el este de África, mientras que A3a sólo se encuentra entre los EA. En el haplogrupo B, B2a y B2b son dos clados hermanos, mientras que B*(xB2) agrega un número de cromosomas de África central que eran ancestrales para el conjunto de SNPs que hemos tipado. B2a tiene una distribución muy amplia y está presente principalmente en las poblaciones bantú-parlantes. Dentro del haplogrupo B2b, B2b* contiene muestras de este, sudeste y centro de África, con P6 derivados de cromosomas procedentes de Sudáfrica y P7 de las poblaciones de CRs procedentes de Europa central, oriental y sur de África. Estos resultados parecen indicar que el sur de África fue el destino inicial de antiguas migraciones humanas procedentes de otras regiones distintas de la fuente original. Con respecto a los roles de África oriental y central, el conjunto de datos que aquí se presenta sigue siendo compatible con una función primordial para el este de África, de acuerdo con las hipótesis generadas a partir del análisis del ADN mitocondrial y el estudio de los primeros fósiles de *Homo sapiens*.

La variabilidad genética en Europa: la península italiana

Implicaciones antropológicas

• La gran importancia del refugio franco-cantábrico para la repoblación de gran parte de Europa Occidental y del Norte a principios del Holoceno, ha sido obtenida por las estimaciones de la edad y distribución geográfica de los haplogrupos del ADNmt, H1, H3, V y U5b1b. Haplogrupos del cromosoma Y, R1b1b2-M269, I1-M253, y M223-I2b1 apoyan el importante papel de la zona de refugio franco-cantábrico, mientras que otros haplogrupos Y (I2a1-M423 y M17-R1a1) revelan que los Balcanes y las zonas de refugio de Ucrania también fueron principales fuentes genéticas para la recolonización humana de Europa. Además de estos refugios, otro refugio glacial en Europa fue la península italiana. Sin embargo, los estudios del ADN mitocondrial y de cromosoma Y han sido capaces de identificar los haplogrupos que marcaron la expansión desde esta área, indicando un papel marginal, en este caso, de esta zona en la repoblación postglacial de Europa.

El haplogrupo U5 es uno de los haplogrupos del ADN mitocondrial más antiguos encontrados en Europa. Se desarrolló principalmente en Europa, donde se extendió después de haber participado en el primer asentamiento del continente para los humanos modernos. Su filogenia se caracteriza por dos ramas U5a y U5b, que son comunes en las poblaciones más europeas, con U5b dividido en U5b1 y U5b2. En 2006, una tercera rama poco común,

llamada U5b3, con el motivo de la región control 16169A-16192-16235-16270-16519-150, se detectó sólo en Cerdeña, una isla que se mantuvo lejana del continente, incluso cuando el nivel del mar era menor durante el LGM y que probablemente fue la última de las grandes islas del Mediterráneo en ser colonizada para los humanos modernos. Haplogrupo U5b3 está prácticamente ausente en el Cercano Oriente y África del Norte y es poco frecuente en Europa, donde, con la excepción del pico de frecuencia en sardos (3,8%), su frecuencia apenas alcanza el 1% en algunas poblaciones del Mediterráneo. Los 55 U5b3 ADNmts detectados en Cerdeña se caracterizan por la región control con motivos del sub-haplogrupo U5b3a1a, cuyo tiempo de coalescencia estimado es entre 4,600-6,300 años. Esto significaría que U5b3a1a surgió *in situ* en Cerdeña después de la llegada de un ADNmt fundador U5b3a1 de algún otro lugar de Europa y que la afiliación U5b3a1a es un marcador de la ascendencia materna de Cerdeña (tiempo de entrada: límite superior es de 9,200-7,200 años (la edad U5b3a1 de nodo), mientras que el límite inferior es 4,600-6,300 años (la edad del nodo U5b3a1a), cuando el subhaplogrupo comenzó a expandirse en Cerdeña. Esta observación preliminar sugiere un vínculo más estrecho entre Cerdeña y el sur de Francia que con otros regiones europeas, incluida Italia continental. Los datos arqueológicos de hace 5.000-10.000 años muestran que la región de Monte Arci del oeste de Cerdeña (provincia de Oristano) fue una de las cuatro fuentes del Mediterráneo (junto con las pequeñas islas de Palmarola, Lipari, y Pantelleria) de la obsidiana, el "oro negro" del Neolítico.

La divergencia del haplogrupo U5b3 es virtualmente idéntica a la descrita para H1 y H3, lo que indica una expansión de la población aproximadamente en el mismo tiempo. Los haplogrupos H1 y H3 difundieron de la zona de refugio franco-cantábrico, cuando mejoraron las condiciones climáticas, por lo tanto, es posible que también la secuencia fundador U5b3 expandió de la misma zona y los tres haplogrupos estuvieron involucrados en los mismos procesos demográficos. Sin embargo, también hay un escenario alternativo: la expansión de U5b3 podría haber ocurrido todavía al mismo tiempo de H1 y H3 cuando las condiciones climáticas en Europa cambiaron, pero a partir de una fuente geográfica distinta. Con la consideración de la distribución moderna de U5b3, el único otro candidato potencial para el último escenario es el refugio glacial en la península italiana.

• Un amplio muestreo a lo largo de la península se hizo para tratar específicamente la cuestión de la distribución de la variación genética del cromosoma Y, a la luz de los modelos

de difusión de la agricultura. Las muestras se distribuyen dentro de la variación genética mostrada por otras poblaciones de Europa y del Mediterráneo. Sin embargo, un cierto grado de separación entre los grupos italianos surgió a lo largo del primer componente principal, con las muestras del norte italiano más cercanas a las poblaciones occidentales de Europa, y las del Sur más cercanas al Sudeste y Sur de los grupos de Europa Central, con pocas excepciones. Las muestras italianas aquí analizadas parecen estar colocadas a lo largo de la clina NO-ES europea, que una serie de estudios anteriores ya han descrito y que se ha considerado como compatible con un escenario demográfico de mezcla entre los agricultores del Cercano Oriente y los habitantes europeos del Mesolítico. Hemos identificado tres clinas para los haplogrupos R1*(xR1a1), J2, E3b1, dos de los cuales mostran también gradientes de diversidad. A la luz del flujo genico procediente de la zona oriental, el análisis de mezcla reveló la introgresión de Anatolia en la mayoría de las muestras italianas. Estos resultados apoyan el modelo de difusión démica (DD). A pesar de la presencia de genes del Neolítico en la actual población italiana masculina, los valores de la mezcla según las estimaciones de ADMIX, sugiere un impacto diferencial de los recién llegados en las muestras italianas. La estimación del grado de introgresión, de hecho, no es coherente en todas las áreas, con las muestras del Sur que experimentan una mayor contribución de Anatolia que las muestras del Norte. Al lado de las contribuciones orientales en las distintas zonas geográficas, el reemplazo de la población no fue completa en toda la península. De ello se deduce que tanto el componentes genético Neolítico como el Mesolíticos se pueden encontrar en el actual pool genético masculino italiano. Es interesante observar que el haplogrupo R1*(xR1a1) muestra una frecuencia clinal opuesta a la encontrada para J2 y E3b1, pero aparentemente sin gradiente de diversidad asociada.

Las poblaciones del Mesolítico tenían baja densidad poblacional y posiblemente un flujo genico limitado a través de los grupos. Si asumimos que la población del Mesolítico se caracteriza por la alta frecuencia del haplogrupo R1*(xR1a1), la variación genética dentro de este haplogrupo sería independiente del muestreo geográfico y en su lugar, principalmente conformada por la historia demográfica local. De ello se deduce que la diversidad de R1*(xR1a1) no se espera que muestre clinas relacionados con la latitud, sino que se distribuya al azar en todas las poblaciones.

Los recién llegados, representados por los agricultores del Neolítico, se han ampliados y mezclados con estos grupos mesolíticos, y esto ha generado, como se esperaba,

clinas de frecuencia y de diversidad a lo largo de la dirección de dispersión, como de hecho se muestra por sus tipos de cromosomas más representativos, E3b1 y J2.

El conjunto actual de datos también proporciona un primer marco para probar la hipótesis de la continuidad genética del Paleolítico al Mesolítico en Italia, a través de la última edad de hielo. Esto apuntaría a la presencia de un refugio del Pleistoceno italiano postulado para Ibéria, Italia y los Balcanes para un número de especies, pero no para los seres humanos.

• Un meta-análisis de los datos de secuencias del cromosoma Y y del ADNmt se han llevado a cabo con el fin de investigar los patrones de variación genética a lo largo de Italia. Índices moleculares indican que la mayoría de las muestras italianas muestran valores de diversidad que son comparables con otras poblaciones europeas. Sin embargo, existen algunas diferencias, especialmente en la población aislada del los Ladinos. Las diferencias regionales son mucho más evidente al examinar las frecuencias de haplogrupos en ambos marcadores uniparentales. Las diferencias son de nuevo más notable para los dos aislados lingüísticos, Ladinos y Grecanos Salentinos.

Los datos muestran que los movimientos demográficos durante el Neolítico han dejado un patrón clinal latitudinal a lo largo de Italia continental. La contribución del Neolítico en la Península Itálica se ha estimado en el cromosoma Y como el 14,5%% y en el ADNmt como el 10,5%. Los datos de cromosoma Y muestran mayor diferenciación entre el Norte, Centro y Sur de la variación del ADNmt. Los AIMs muestran que la variación de los italianos encaja muy bien con otras poblaciones europeas, y aunque estos marcadores no fueron eligidos para permitir una resolución de grano fino de la variación intercontinental, todavía reconocen la presencia de una mínima aportación procedente de la región subsahariana. Esta contribución del sur del Sahara también se detecta en los marcadores uniparentales. El análisis de mezclas basado en el metodo Bayesiano sobre los datos del ADN mitocondrial, lleva a la conclusión preliminar de que el norte de África contribuyó con un total de 27% de la variación en Italia, mientras que el Oriente Medio contribuyó alrededor del 28%.

Italia muestra los patrones de variación molecular que son similares a los de otros países europeos, aunque existe cierta heterogeneidad según se infiere de diferentes análisis y

marcadores moleculares. De Norte a Sur, Italia muestra los patrones de clina modulados durante el Neolítico más probables.

• Un análisis más amplio para determinar la epoca de la distribución espacial del haplogrupo de cromosoma Y, R-M269, que se puede dividir por R-S127 en linajes de Europa y linajes occidentales de Eurasia, se llevó a cabo.

No vimos ninguna relación entre la diversidad y la longitud para el R-M269. Esta correlación es el principio central de la hipótesis de que el R-M269 se extendió con la expansión de los agricultores neolíticos.

La datacion de los linajes de cromosoma Y es notoriamente controvertida: la elección de la tasa de mutación de los STRs puede dar lugar a estimaciones de la edad que difieren por un factor de tres. Al tratar de encontrar un conjunto adecuado de STRs al fin de estimar el tiempo promedio de coalescencia (T) del sub-haplogrupo R-S127, hemos demostrado que no todos los STRs son de igual uso en este contexto. Nos concentramos en la estimación de la duración de la linealidad (D), utilizando diferentes conjuntos de STRs. Nuestros análisis sugieren que la D de una STR es la clave para su capacidad de descubrir ascendencia remota.

• La información combinada de los datos de secuencias de mtSNPs y HVS-I, nos confirmó que alrededor del 10% (26 personas de 258) de los toscanos en realidad pertenecen a una de los típicos haplogrupos del Cercano Este, con también una correspondencia con poblaciones de esta región geográfica. Todos estos haplotipos de Oriente Cercano son diversos (con la excepción de aquellos que pertenecen a U7) y caen en la punta de la filogenia, lo que sugiere un reciente llegada a la región. El típico haplogrupo de Oriente Próximo, U7, ocurre con frecuencia relativamente alta en la Isla de Elba (~ 17%, 9 ADNmt de 53), y todos estos U7-ADNmts comparten el mismo motivo de HVS-I (T16271C-A16318T-T16519C), lo que indica que este linaje podría representar uno de los fundadores de Cercano Oriente en la isla. La secuenciación completa del genoma de estos nueve U7-ADNmts permitió la identificación de un nuevo subclade, U7a2a, que se caracteriza por las transiciones A13395G y T16271C. U7a2a es una sub-rama de U7a2. La cantidad de variación acumulada en el grupo U7a2a etrusco (suponiendo un único fundador) puede fecharse en el rango de 1,1 ± 0,1 a 2,3 ± 0,4 kya, de acuerdo con una reciente llegada de este haplogrupo a esta isla y compatibles con la cultura etrusca (9º-1er siglo a.C.).

• Para estimar al grado de contribución histórica de Africa Noroeste (NO) en el Mediterráneo, se utilizó una combinación de Y-STRs/SNPs. Los tiempos de coalescencia para los tres haplogrupos específicos NO africanos se situan entre 5.000 y 24.000 años, abarcando una serie de escenarios históricos que pueden cada uno potencialmente explicar esta presencia en las costas del norte del Mediterráneo. De esto se deduce que la estimación de la herencia genética Medieval del Norte de África (MNA), sobre sólo la base de haplogrupos, sería un error. Para evitar esta limitación, hemos ampliado nuestro análisis para incluir los datos de STRs, que por su alta tasa de mutación nos permite centrar en los acontecimientos más recientes. Hemos examinado más de 2300 muestras del sur de Europa para identificar a los haplotipos que son evolutivamente cerca de cromosomas NO africanos. Las frecuencias totales de estos cromosomas estan entre 0 y 19% en el sur de Europa, siendo más elevada en Cantabria y incluyendo una muestra del Valle del Pas.

Nuestras estimaciones de frecuencias de cromosomas NO africanos fueron más altas en la Península Ibérica y Sicilia, de acuerdo con la dominación árabe que hubo en estas dos áreas. Las frecuencias de los cromosomas en las dos muestras no fueron significativamente diferentes entre sí (prueba exacta de Fisher, P = 0,83), pero fueron significativamente diferentes de las muestras de la península Italiana (P = 0,01). Haplotipos MNA en la península italiana se encontraron en tres muestras geográficamente cercanas en todo el sur de los Apeninos (este de Campania, Apulia noroeste, Lucera). Cuando se agruparon, estas tres muestras italianas indicaron una frecuencia local del 4,7%, significativamente diferente de las del Norte y del resto del sur de Italia (P = 0,01), pero no de la Península Ibérica y de Sicilia (P = 0,12 yp = 0,33, respectivamente). La presencia árabe se registra históricamente en estas áreas después de la reubicación de Federico II de los árabes de Sicilia. En la península Ibérica, una distribución no aleatoria también puede potencialmente estar presente, según lo sugerido por nuestras estimaciones más bajas en el noreste (País Vasco y catalanes), pero un muestreo mas amplio de toda la península seria ideal para abordar adecuadamente esta cuestión.

Implicaciones forenses

Cromosoma Y

• Un total de 186 haplotipos fueron definidos a través de comparaciones haplotipicas de un conjunto de 9 Y-STRs. Ciento treinta y cinco haplotipos de nuestra colección (72,5% del total) no estaban presentes en la base de datos anterior. Este número de nuevos tipos, definidos por nuestros marcadores, sugiere que la diversidad de cromosoma Y en Italia está sólo parcialmente conocida y que una considerable heterogeneidad existe. Un muestreo adicional y extenso sería necesario para obtener una descripción más detallada de la composición haplotipica de la Península.

Comparaciones se realizaron también utilizando los datos publicados disponibles de Europea y de las poblaciones del norte de África. El análisis de AMOVA reveló una variación genética significativa en las tres áreas italianas (porcentaje de la variación genética entre las poblaciones: 1,25, $p < 0,05$). Ninguna diferenciación significativa de la población se encontró con la prueba exacta de Fisher. Una variación significativa se encontró cuando las tres muestras se agruparon y se testaron con la base de datos GEFI (AMOVA, el porcentaje de la variación genética entre las poblaciones: 0,24, $p < 0,05$).

• El objetivo de otro estudio sobre cromosoma Y era contribuir a una base de datos italiana de interés forense, dado que: (i) la mayor parte de los datos de cromosoma Y disponibles en literatura se limita en el número de loci (haplotipo mínimo), (ii) en general, se centran en poblaciones no representativas, o (iii) cubren la península italiana sólo parcialmente. Nuestro muestreo se constituye por individuos del Norte, Centro y Sur de Italia y ha sido genotipo para el conjunto completo de Y-STRs incluidos en el kit de amplificación AmpFlSTR Yfiler (AB). Hemos demostrado que el análisis de perfiles completos de Y-STR (Yfiler) vs haplotipos mínimos es importante para la detección de diferencias genéticas locales, y que la discriminación genética proporcionada para los perfiles Yfiler es sustancialmente mayor que para el haplotipo mínimo.

El análisis de AMOVA demonstrò que la medida de estratificación de la población no sólo depende de la cantidad de información genética usada, sino también, del numero de muestras de poblaciónes utilizadas en el análisis. Esto garantiza una mayor precaución al usar los valores de FST para corregir la estratificación de la población en casos forenses.

• Las duplicaciónes del locus DYS19 existen en todas las poblaciones. Individuos asiáticos que llevan esta duplicación pertenecen al haplogrupo C3c. Diez nuevos duplicaciónes del DYS19 han sidos identificadas: todos estos cromosomas partenecen al haplogrupo G2*(xG2a, b). Nuestro hallazgo implica claramente que la duplicación en el DYS19 tuvo que ocurrir al menos dos veces, impulsada por los muchos eventos genéticos independientes. A la luz de la existencia actual de los dos linajes, la hipótesis del origen único implica que la duplicación original tuvo que ocurrir muy temprano en la escala de tiempo de la evolución humana. En el caso de un solo origen, eventos múltiples de reversión deberían haber ocurrido. La hipótesis del origen múltiple, necesitando un mínimo de sólo dos eventos de duplicación independientes, aparece la más probable y más parsimoniosa .

Datos Autosómico

• Quince STRs autosómicos incluidos en el sistema multiplex PowerPlex16 (Promega Corporation, Madison, WI) se analizaron en más de 400 muestras de 9 diferentes zonas de Italia.

No se detectaron diferencias significativas entre las poblaciones. La prueba se llevó a cabo teniendo en cuenta el AMOVA para cada población de forma independiente o agrupada de acuerdo a su origen geográfico (Norte, Centro y Sur de Italia). Los resultados mostraron que la mayor parte de la variación genética se produce en los individuos, lo que subraya que no hay evidencia de estructura en la población. El poder combinado de exclusión (PE) y el poder de discriminación (PD) para los 15 loci estudiados fueron 0,964708775 y 0,9999999997, respectivamente. El valor de probabilidad combinada fue de 1 en 3.33x108. Sobre la base de heterocigosidad y contenido de información polimórfica (PIC), FGA se puede considerar como el locus más informativo. Las frecuencias alélicas de cada locus se compararon con los datos publicados anteriormente de la población italiana y mediterránea. Después de aplicar la corrección de Bonferroni para pruebas múltiples, las pruebas de diferenciación de la población mostró que Italia tiene diferencias significativas con Bosnia-Herzegovina en 4 de los 13 loci (D5S818, D7S820, D16S539 y CSF1PO), con España en 3 de los 13 loci (D3S1358, TH01 y D16S539) y con los albaneses de Kosovo en 1 de cada 13 loci (D18S51).

• Las frecuencias allelicas de los 5 nuevos loci STRs, incluidas en el ESI PowerPlex ® 17 System (Promega Corporation, Madison, WI) se calcularon en una muestra de 209 italianos no relacionados.

El poder combinado de exclusión (PE) y el poder de discriminación (PD) para los 16 loci estudiados fueron 0,999999935 y 0,999999999, respectivamente. Sobre la base de la heterocigosidad y contenido de información polimórfica (PIC), SE33 se puede considerar el locus loci más informativo. La exclusión de este locus reduce ligeramente la estimación del PE (0,999999555). El valor de la PE es similar a los valores calculados en una población italiana diferente mediante el sistema Multiplex PowerPlex16 (Promega Corporation, Madison, WI), el kit de Identifiler1 (Applied Biosystems) y de los marcadores incluidos en el Sistema de Índice Combinado de ADN EE.UU. (CODIS) , mientras que el valor obtenido con el ESI PowerPlex ® 17 System (Promega Corporation, Madison, WI) es mayor.

ADN antiguo.

Afiliación filogenética del Mamut.

• La filogenia de los elefántidos no puede ser resuelta con el conjunto de datos actuales.

Sin embargo, la tendencia no sugiere una fuerte asociación de Loxodonta-Mamut como ha sido reportado en varios estudios basados en ADN mitocondrial. Al contrario, el gen VWF sugiere una asociación Mamut-Elephas, como lo hace el gen BGN. Mientras que el GBA es ambiguo, CHRNA1 favorece una asociación de Mamut-Loxodonta y LEPR favorece ligeramente Loxodonta-Mamut, ya que hay individuos heterocigotos de elefantes de foresta con una sola diferencia en comparación con el mamut, mientras que los elefantes asiáticos de manera uniforme mostran dos diferencias LEPR fijas, frente a la secuencia de mamut. Sin embargo, ninguno de nuestros análisis de la combinación de todas las secuencias detectó una agrupación Mamut-Loxodonta.

Conclusiones

África subsahariana

Camerún

1) El efecto combinado de la propensión a la mezcla inter-poblacional de las mujeres, favorecida por los contactos culturales (en especial los factores lingüísticos), y de la deriva genética que actúa sobre la diversidad del cromosoma Y, podría explicar el patrón genético

peculiar observado en el norte de Camerún. Nuestros resultados están en línea con la hipótesis de una tendencia hacia el aislamiento entre los *Montagnard*s del norte de Camerún, en relación con las barreras físicas creadas por el entorno montañoso.

África occidental y meridional

2) La distribución de los haplogrupos A y B esta casi exclusivamente restringida a África subsahariana, con un pico de 65% en el grupo de cazadores-recolectores. Linajes específicos relacionados con la dinámica regional preagrícolas en diferentes áreas del África subsahariana, contribuyen a entender la compleja relación evolutiva entre los cazadores-recolectores de África (los pigmeos y los grupos Koisan hablantes del sur de África). El papel del sur de África puede ahora ser visto más como disipador que de fuente de las primeras migraciones de los humanos modernos de la parte oriental y central del continente.

La variabilidad genética en Europa: visión de la península italiana

Conclusiones antropológicas

3) La expansión del Holoceno (~ 11 kya) del haplogrupo U5b3 ocurrió a lo largo de las costas mediterráneas, principalmente hacia la cercana Provenza (sur de Francia). A partir de ahí, ~ 7.000-9.000 años atrás, este subclade se trasladó a Cerdeña (comercio de obsidiana), dejando una firma distintiva en la gente moderna de la isla. Este escenario coincide sorprendentemente con la edad, la distribución y el origen geográfico postulado de un haplogrupo de cromosoma Y de Cerdeña (I2a2-M26), un caso paradigmático en el contexto europeo de un evento fundador de marcar los dos linajes femeninos y masculinos.

4) El papel de los agricultores neolíticos fue mayor que el de los históricos colonizadores griegos del sur de Italia, y el resultado apoya un modelo de difusión demica masculina (la diversidad distribuidos a lo largo de un eje Norte-Sur), aunque el reemplazo de la población no fue completo y el grado de mezcla neolítica con los habitantes del mesolítico fue diferente en diferentes zonas de Italia; la deriva definitivamente tuvo un papel en la conformación de la variación genética del actual cromosoma Y.

5) Italia muestra clinas de variación atribuibles a los movimientos demográficos de los primeros asentamientos del Paleolítico, posteriormente modelados por el Mesolítico y por, en medida menor, los agricultores neolíticos. Las diferencias regionales surgieron con el

tiempo, y son más notables en los aislados lingüísticos, como los ladinos, y en manera minor, en los Grecanos Salentinos

6) La opinión sobre el origen del haplogrupo más importantes de Europa de cromosoma Y, R-M269, se ha alejado de los orígenes en el Paleolítico, a la noción de una recien propagación Neolítica de este cromosoma desde el Cercano Oriente No se observaron tendencias geográficas en la diversidad, en contradicción con las expectativas en la hipótesis del Neolítico, y esto sugiere una explicación alternativa para la clina de diversidad de reciente descripción. La frecuencia de R-M269 en toda Europa podría estar relacionada con el crecimiento de múltiples, geográficos y específicos sublinajes que difieren en diferentes partes de Europa.

7) La investigación de una serie amplia y representativa de muestras y el análisis de genomas completos del ADN mitocondrial apoyan la hipótesis de que la Toscana aún conserva la huella de una conexión histórica con el Cercano Oriente. Sin embargo, cabe destacar que esto representa sólo una componente menor de la estructura genética toscana y sugiere que históricamente las diferentes capas se superponen al pool genético del Mesolítico de la Península.

8) Más del 56% de los individuos italianos aquí identificados, teniendo una reciente introgresión Noroeste (NO) africana, no tienen "match" en la base de datos italiana de cromosoma Y (1200 individuos). De éstos, 31% coinciden perfectamente con los tipos de poblaciones NO africanas. Nuestros resultados confirman claramente que las conclusiones basadas en los cromosomas uniparentales se deben tomar con mucha cautela.

Conclusiones forenses

9) Un gran número de nuevos tipos sugiere que la diversidad de cromosoma Y definida por nuestro conjunto de marcadores (9 Y-STRs) esta en Italia sólo parcialmente conocida y que una considerable heterogeneidad existe. Un muestreo adicional y extenso, sería necesario para obtener una descripción más detallada de la composición haplotipica de la Península.

10) Aunque Italia parece mostrar niveles más bajos de varianza entre las poblaciónes que otras regiones europeas, estos valores son lo suficientemente altos (alrededor de 3-11%) para garantizar el cuidado al utilizar la prueba del cromosoma Y en la ciencia forense. Por lo

tanto, los resultados plantean la cuestión del desarrollo de bases de datos locales de referencia en Italia para fines forenses, dado las diferencias sustanciales que existen en la península italiana

11) El análisis de redes basado en los haplogrupo y microsatélites de cromosoma Y, parecen indicar que las duplicaciones de DYS19 no tienen un origen comun, sino que ocurrieron más de una vez durante la evolución humana, en concreto por no-recombinación homóloga alélica o por conversión de genes

12) Una serie de bases de datos autosómicos italianos están actualmente disponibles. Sin embargo, estan restringidos en el número de loci, se centran en poblaciones no representativas o cubren la península italiana sólo parcialmente. Nuestro muestreo comprende individuos de Norte, Centro y Sur de Italia, y el análisis de 15 loci (16 Powerplex sistema múltiplex, Promega Corporation, Madison, WI), proporciona información adicional sobre la variación genética de la población italiana.

13) Los índices forenses de los cinco nuevos STRs autosómicos incluidos en el sistema ESI Powerplex® 17 (Promega Corporation, Madison, WI) se calcularon. Las comparaciones de estos resultados con los de otros kits comerciales (PowerPlex16 sistema múltiplex, Identifiler kit, EE.UU. Sistema de Índice Combinado de ADN (CODIS), indica que, en función de los STRs, los valores de los distintos índices utilizados en las pruebas de paternidad o en las causas penales pueden variar.

Mamut y su afiliación filogenética

14) Los resultados del estudio sugieren que el uso de los genes nucleares de mamut lanudo debería resolver su filogenia de manera concluyente. Los recientes acontecimientos en la tecnología de secuenciación sugieren que el análisis del genoma completo de los animales extintos, los mamuts en particular, será factible. También se concluye que la aplicación de marcadores nucleares es ahora posible y, de hecho preferible para el estudio sistemático de una amplia variedad de animales extintos, representados por restos bien conservados en las colecciónes de museos.

Bibliography

7. Bibligraphy

Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, Battaglia V, Accetturo M, Kutuev I, Khusnutdinova E, Pennarun E, Cerutti N, Di Gaetano C, Crobu F, Palli D, Matullo G, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Semino O, Villems R, Bandelt H-J, Piazza A, Torroni A (2007) Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. Am. J. Hum. Genet. 80: 759-68

Achilli A, Olivieri A, Pellecchia M, Uboldi C, Colli L, Al-Zahery N, Accetturo M, Pala M, Kashani BH, Perego UA, Battaglia V, Fornarino S, Kalamati J, Houshmand M, Negrini R, Semino O, Richards M, Macaulay V, Ferretti L, Bandelt HJ, Ajmone-Marsan P, Torroni A (2008) Mitochondrial genomes of extinct aurochs survive in domestic cattle. Curr Biol 18: R157-8

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1994) Molecular Biology of the Cell, 3rd edition. New York: Garland Science.

Álvarez-Iglesias V, Jaime JC, Carracedo A, Salas A (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int Genet 1: 44-55

Alves-Silva J, da Silva Santos M, Guimarães PE, Ferreira AC, Bandelt HJ, Pena SD, Prado VF (2000) The ancestry of Brazilian mtDNA lineages. Am J Hum Genet 67: 444-61

Amigo J, Phillips C, Salas A, Carracedo A (2009) Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes. BMC Bioinformatics 10: S5

Amigo J, Salas A, Phillips C, Carracedo Á (2008) SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. BMC Bioinformatics 9: 428

Ammerman AJ, Cavalli-Sforza LL (1984) The Neolithic transition and the genetics of populations in Europe.

Anderson EC, Garza JC (2006) The power of single nucleotide polymorphisms for large-scale parentage inference. Genetics 172: 2567-2582

Asamura H, Sakai H, Ota M, Fukushima H (2007) MiniY-STR quadruplex systems with short amplicon lengths for analysis of degraded DNA samples. Forensic Sci Int Genet 1: 56-61

Ayub Q, Mohyuddin A, Qamar R, Mazhar K, Zerjal T, Mehdi SQ, Tyler-Smith C (2000) Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. Nucleic Acids Res 28: e8

Babalini C, Martínez-Labarga C, Tolk HV, Kivisild T, Giampaolo R, Tarsi T, Contini I, Barać L, Janićijević B, Martinović Klarić I, Peričić M, Sujoldžić A, Villems R, Biondi G, Rudan P, Rickards O (2005) The population history of the Croatian linguistic minority of Molise (southern Italy): a maternal view. Eur J. Hum. Genet. 13: 902-12

Bandelt HJ (2004) Etruscan artifacts. Am J Hum Genet 75: 919-920; author reply 923-927

Bandelt HJ, Kong QP, Parson W, Salas A (2005) More evidence for non-maternal inheritance of mitochondrial DNA? J Med Genet 42: 957-60

Bandelt HJ, Lahermo P, Richards M, Macaulay V (2001) Detecting errors in mtDNA data by phylogenetic analysis. Int J Legal Med 115: 64-9

Bandelt HJ, Salas A, Bravi C (2004a) Problems in FBI mtDNA database. Science 305: 1402-4

Bandelt HJ, Salas A, Lutz-Bonengel S (2004b) Artificial recombination in forensic mtDNA population databases. Int J Legal Med 118: 267-73

Barnes I, Matheus P, Shapiro B, Jensen D, Cooper A (2002) Dynamics of Pleistocene population extinctions in Beringian brown bears. Science 295: 2267-70

Bässler G, Förster R, Eberspächer B, Karl C, Kugler M, Pflug W (1999) Frequency data for the STR loci HumFibra (FGA) and HumACTBP2 (SE33) in a population of Germans and Turks from South-West Germany. Int J Legal Med 112: 136-8

Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, Comas D, Destro-Bisol G, Calafell F (2007) Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. Mol Phylogenet Evol 43: 635-44

Batini C, Ferri G, Destro-Bisol G, Brisighelli F, Luiselli D, Sánchez-Diz P, Rocha J, Simonson T, Brehm A, Montano V, Elwali NE, Spedini G, D'Amato ME, Myres N, Ebbesen P, Comas D, Capelli C (2011) Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. Mol Biol Evol 28: 2603-13

Battaglia V, Fornarino S, Al-Zahery N, Olivieri A, Pala M, Myres NM, King RJ, Rootsi S, Marjanovic D, Primorac D, Hadziselimovic R, Vidovic S, Drobnic K, Durmishi N, Torroni A, Santachiara-Benerecetti AS, Underhill PA, Semino O (2009) Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. Eur J Hum Genet 17: 820-30

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S (2008) The dawn of human matrilineal diversity. Am J Hum Genet 82: 1130-40

Beleza S, Gusmão L, Amorim A, Carracedo A, Salas A (2005) The genetic legacy of western Bantu migrations. Hum Genet 117: 366-75

Bendall KE, Macaulay VA, Baker JR, Sykes BC (1996) Heteroplasmic point mutations in the human mtDNA control region. Am J Hum Genet 59: 1276-87

Bianchi NO, Catanesi CI, Bailliet G, Martinez-Marignac VL, Bravi CM, Vidal-Rioja LB, Herrera RJ, López-Camelo JS (1998) Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. Am J Hum Genet 63: 1862-71

Bini C, Ceccardi S, Luiselli D, Ferri G, Pelotti S, Colalongo C, Falconi M, Pappalardo G (2003) Different informativeness of the three hypervariable mitochondrial DNA regions in the population of Bologna (Italy). Forensic Sci. Int. 135: 48-52

Blockely SPE, Pinhasi R (2011) A revised chronology for the adoption of agriculture in the Southern Levant and the role of the Lateglacial climatic change. Q Sci Rev 30: 98-108

Bortolini MC, Salzano FM, Thomas MG, Stuart S, Nasanen SP, Bau CH, Hutz MH, Layrisse Z, Petzl-Erler ML, Tsuneto LT, Hill K, Hurtado AM, Castro-de-Guerra D, Torres MM, Groot H, Michalski R, Nymadawa P, Bedoya G, Bradman N, Labuda D, Ruiz-Linares A (2003) Y-chromosome evidence for differing ancient demographic histories in the Americas. Am J Hum Genet 73: 524-39

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, Antanaitis-Jacobs I, Haidle MN, Jankauskas R, Kind CJ, Lueth F, Terberger T, Hiller J, Matsumura S, Forster P, Burger J (2009) Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science 326: 137-40

Brandstätter A, Salas A, Niederstätter H, Gassner C, Carracedo A, Parson W (2006) Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 27: 2541-50

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62: 1408-15

Brión M, Sánchez JJ, Balogh K, Thacker C, Blanco-Verea A, Børsting C, Stradmann-Bellinghausen B, Bogus M, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N (2005a) Introduction of an single nucleodite polymorphism-based "Major Y-chromosome haplogroup typing kit" suitable for predicting the geographical origin of male lineages. Electrophoresis 26: 4411-4420

Brión M, Sobrino B, Blanco-Verea A, Lareu MV, Carracedo Á (2005b) Hierarchical analysis of 30 Y-chromosome SNPs in European populations. Int J Legal Med 119: 10-5

Brisighelli F, Capelli C, Álvarez-Iglesias V, Onofri V, Paoli G, Tofanelli S, Carracedo Á, Pascali VL, Salas A (2009) The Etruscan timeline: a recent Anatolian connection. Eur J. Hum. Genet. 17: 693-6

Brown WM, George M, Jr., Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci U S A 76: 1967-71

Burton ML, Moore CC, Whiting JWM, Romney AK (1996) Regions based on social structure. Current Anthropology 37: 87-123

Busby GB, Brisighelli F, Sánchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, Thomas MG, Bradley DG, Gusmão L, Winney B, Bodmer W, Vennemann M, Coia V, Scarnicci F, Tofanelli S, Vona G, Ploski R, Vecchiotti C, Zemunik T, Rudan I, Karachanak S, Toncheva D, Anagnostou P, Ferri G, Rapone C, Hervig T, Moen T, Wilson JF, Capelli C (2011) The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc Biol Sci

Butler JM, Shen Y, McCord BR (2003) The development of reduced size STR amplicons as tools for analysis of degraded DNA. J Forensic Sci 48: 1054-64

Caglià A, Tofanelli S, Coia V, Boschi I, Pescarmona M, Spedini G, Pascali V, Paoli G, Destro-Bisol G (2003) A study of Y-chromosome microsatellite variation in sub-Saharan Africa: a comparison between F(ST) and R(ST) genetic distances. Hum Biol 75: 313-30

Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK (1998) Short tandem repeat polymorphism evolution in humans. Eur J Hum Genet 6: 38-49

Cali F, Le Roux MG, D'Anna R, Flugy A, De Leo G, Chiavetta V, Ayala GF, Romano V (2001) MtDNA control region and RFLP data for Sicily and France. Int J Legal Med 114: 229-31

Campbell MC, Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9: 403-33

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325: 31-6

Capelli C, Brisighelli F, Scarnicci F, Arredi B, Caglia A, Vetrugno G, Tofanelli S, Onofri V, Tagliabracci A, Paoli G, Pascali VL (2007) Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. Mol Phylogenet Evol 44: 228-39

Capelli C, Onofri V, Brisighelli F, Boschi I, Scarnicci F, Masullo M, Ferri G, Tofanelli S, Tagliabracci A, Gusmão L, Amorim A, Gatto F, Kirin M, Merlitti D, Brión M, Verea AB, Romano V, Cali F, Pascali V (2009) Moors and Saracens in Europe: estimating the medieval North African male legacy in southern Europe. Eur J Hum Genet 17: 848-52

Capelli C, Redhead N, Abernethy JK, Gratrix F, Wilson JF, Moen T, Hervig T, Richards M, Stumpf MP, Underhill PA, Bradshaw P, Shaha A, Thomas MG, Bradman N, Goldstein DB (2003) A Y chromosome census of the British Isles. Curr Biol 13: 979-84

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 (2006) Population structure in the Mediterranean basin: a Y chromosome perspective. Ann Hum Genet 70: 207-25

Capelli C, Wilson JF, Richards M, Stumpf MP, Gratrix F, Oppenheimer S, Underhill P, Pascali VL, Ko TM, Goldstein DB (2001) A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. Am J Hum Genet 68: 432-43

Casanova M, Leroy P, Boucekkine C, Weissenbach J, Bishop C, Fellous M, Purrello M, Fiori G, Siniscalco M (1985) A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science 230: 1403-6

Cavalli-Sforza L, Menozzi P, Piazza A (1994) The history and geography of human genes. Charter 2. Princeton, New Jersey: Princeton University Press.

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19: 223-257

Cerný V, Hájek M, Cmejla R, Brůzek J, Brdicka R (2004) mtDNA sequences of Chadic-speaking populations from northern Cameroon suggest their affinities with eastern Africa. Ann Hum Biol 31: 554-69

Cerný V, Salas A, Hájek M, Zaloudková M, Brdicka R (2007) A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet 71: 433-452

Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B (1999) The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis 20: 1682-96

Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC (1995) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet 57: 133-49

Chikhi L, Nichols RA, Barbujani G, Beaumont MA (2002) Y genetic data support the Neolithic demic diffusion model. Proc Natl Acad Sci U S A 99: 110-118

Cinnioğlu C, King R, Kivisild T, Kalfoğlu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA (2004) Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet 114: 127-48

Coble MD, Butler JM (2005) Characterization of new miniSTR loci to aid analysis of degraded DNA. J Forensic Sci 50: 43-53

Coia V, Brisighelli F, Donati F, Pascali V, Boschi I, Luiselli D, Battaggia C, Batini C, Taglioli L, Cruciani F, Paoli G, Capelli C, Spedini G, Destro-Bisol G (2009) A multi-perspective view of genetic variation in Cameroon. Am J Phys Anthropol 140: 454-64

Coia V, Caglià A, Arredi B, Donati F, Santos FR, Pandya A, Taglioli L, Paoli G, Pascali V, Spedini G, Destro-Bisol G, Tyler-Smith C (2004) Binary and microsatellite polymorphisms of the Y-chromosome in the Mbenzele pygmies from the Central African Republic. Am J Hum Biol 16: 57-67

Coia V, Destro-Bisol G, Verginelli F, Battaggia C, Boschi I, Cruciani F, Spedini G, Comas D, Calafell F (2005) Brief communication: mtDNA variation in North Cameroon: lack of

Asian lineages and implications for back migration from Asia to sub-Saharan Africa. Am J Phys Anthropol 128: 678-81

Collard M, Edinborough K, Shennan S, Thomas MG (2010) Radiocarbon evidence indicates that migrants introduced farming to Britain. J Archaeol Sci 37: 866-870

Contu D, Morelli L, Santoni F, Foster JW, Francalacci P, Cucca F (2008) Y-chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans. PLoS One 3: e1430

Cooper G, Amos W, Hoffman D, Rubinsztein DC (1996) Network analysis of human Y microsatellite haplotypes. Hum Mol Genet 5: 1759-66

Cox MP, Mirazón Lahr M (2006) Y-chromosome diversity is inversely associated with language affiliation in paired Austronesian- and Papuan-speaking communities from Solomon Islands. Am J Hum Biol 18: 35-50

Cruciani F, La Fratta R, Santolamazza P, Sellitto D, Pascone R, Moral P, Watson E, Guida V, Colomb EB, Zaharova B, Lavinha J, Vona G, Aman R, Cali F, Akar N, Richards M, Torroni A, Novelletto A, Scozzari R (2004) Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. Am J Hum Genet 74: 1014-22

Cruciani F, La Fratta R, Torroni A, Underhill PA, Scozzari R (2006) Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers. Hum Mutat 27: 831-2

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 (2007) Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. Mol Biol Evol 24: 1300-11

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet 70: 1197-214

Cunliffe B (2001) The Oxford Illustrated History of Prehistoric Europe. .

David N (1981) The archaeological background of Cameroonian history. In: Tardits C, editor. Colloques internationales du C.N.R.S. Paris. 551: 80-97

Davies RHC (1988) A History of Medieval Europe. 83-101

Deka R, Jin L, Shriver MD, Yu LM, DeCroo S, Hundrieser J, Bunker CH, Ferrell RE, Chakraborty R (1995) Population genetics of dinucleotide (dC-dA)n.(dG-dT)n polymorphisms in world populations. Am J Hum Genet 56: 461-74

Destro-Bisol G, Coia V, Boschi I, Verginelli F, Caglià A, Pascali V, Spedini G, Calafell F (2004) The analysis of variation of mtDNA hypervariable region 1 suggests that Eastern and Western Pygmies diverged before the Bantu expansion. Am Nat 163: 212-26

Di Giacomo F, Luca F, Popa LO, Akar N, Anagnou N, Banyko J, Brdicka R, Barbujani G, Papola F, Ciavarella G, Cucci F, Di Stasi L, Gavrila L, Kerimova MG, Kovatchev D, Kozlov AI, Loutradis A, Mandarino V, Mammì C, Michalodimitrakis EN, Paoli G, Pappa KJ, Pedicini G, Terrenato L, Tofanelli S, Malaspina P, Novelletto A (2004) Y

chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet 115: 357-371

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

Di Rienzo A, Wilson AC (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc. Natl. Acad. Sci. U S A. 88: 1597-1601

Diamond J, Bellwood P (2003) Farmers and their languages: the first expansions. Science 300: 597-603

Dupuy BM, Stenersen M, Egeland T, Olaisen B (2004) Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. Hum Mutat 23: 117-24

Ehret C (1984) Historical/linguistic evidence for early African food production. In: Clark JD, Brandt SA, editors. From hunters to farmers. Berkeley: University of California Press.: 26-35

Evett IW, Weir BS (1998) Interpreting DNA evidence, Statistical Genetics for dorensic scientists. Sinauer Associates Inc.

Falchi A, Giovannoni L, Calo CM, Piras IS, Moral P, Paoli G, Vona G, Varesi L (2006) Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms. J. Hum. Genet. 51: 9-14

Ferri G, Alu M, Corradini B, Radheshi E, Beduschi G (2009) Slow and fast evolving markers typing in Modena males (North Italy). Forensic Sci Int Genet 3: e31-3

Ferri G, Ceccardi S, Lugaresi F, Bini C, Ingravallo F, Cicognani A, Falconi M, Pelotti S (2008) Male haplotypes and haplogroups differences between urban (Rimini) and rural area (Valmarecchia) in Romagna region (North Italy). Forensic Sci Int 175: 250-5

Finnilä S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. Am J Hum Genet 68: 1475-84

Forster P, Cali F, Rohl A, Metspalu E, D'Anna R, Mirisola M, De Leo G, Flugy A, Salerno A, Ayala G, Kouvatsi A, Villems R, Romano V (2002) Continental and subcontinental distributions of mtDNA control region types. Int. J. Legal Med. 116: 99-108

Forster P, Harding R, Torroni A, Bandelt HJ (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59: 935-45

Forster P, Röhl A, Lünnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, Brinkmann B (2000) A short tandem repeat-based phylogeny for the human Y chromosome. Am J Hum Genet 67: 182-96

Francalacci P, Bertranpetit J, Calafell F, Underhill PA (1996) Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. Am J Phys Anthropol 100: 443-460

Fung WK, Chung YK, Wong DM (2002) Power of exclusion revisited: probability of excluding relatives of the true father from paternity. Int J Legal Med 116: 64-7

Gallagher A, Gunther MM, Bruchhaus H (2009) Population continuity, demic diffusion and Neolithic origins in central-southern Germany: the evidence from body proportions. Homo 60: 95-126

Gamble C, Davies W, Pettitt P, Hazelwood L, Richards M (2005) The archaeological and genetic foundations of the European population during the Late Glacial: implications for 'agricultural thinking'. Camb Archaeol J 15: 193-223

Garrigan D, Hammer MF (2006) Reconstructing human origins in the genomic era. Nat Rev Genet 7: 669-80

Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci U S A 77: 6715-9

Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. Int J Legal Med 114: 204-10

Gill P, Brenner C, Brinkmann B, Budowle B, Carracedo A, Jobling MA, de Knijff P, Kayser M, Krawczak M, Mayr WR, Morling N, Olaisen B, Pascali V, Prinz M, Roewer L, Schneider PM, Sajantila A, Tyler-Smith C (2001) DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. Forensic Sci Int 124: 5-10

Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, Morling N, Prinz M, Schneider PM, Weir BS (2006) DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Sci Int 160: 90-101

Goebel T (2007) Anthropology. The missing years for modern humans. Science 315: 194-6

Gómez J, Carracedo A (2000) The 1998-1999 collaborative exercises and proficiency testing program on DNA typing of the Spanish and Portuguese Working Group of the International Society for Forensic Genetics (GEP-ISFG). Forensic Sci Int 114: 21-30

Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 24: 757-68

González-Neira A, Elmoznino M, Lareu MV, Sánchez-Diz P, Gusmão L, Prinz M, Carracedo A (2001) Sequence structure of 12 novel Y chromosome microsatellites and PCR amplification strategies. Forensic Sci Int 122: 19-26

Greenberg J (1963) The languages of Africa. Bloomington: Indiana University Publications.

Grubweisser P, Mühlmann R, Berger B, Niederstätter A, Pavlic M, Parson W (2006) A new "miniSTR-multiplex" displaying reduced amplicon lengths for the analysis of degraded DNA. Int J Leg Med 120: 115-120

Gusmão L, Butler JM, Carracedo A, Gill P, Kayser M, Mayr WR, Morling N, Prinz M, Roewer L, Tyler-Smith C, Schneider PM (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. Int J Legal Med 120: 191-200

Gusmão L, González-Neira A, Alves C, Lareu M, Costa S, Amorim A, Carracedo A (2002) Chimpanzee homologous of human Y specific STRs. A comparative study and a proposal for nomenclature. Forensic Sci Int 126: 129-36

Gusmão L, Sánchez-Diz P, Calafell F, Martín P, Alonso CA, Álvarez-Fernández F, Alves C, Borjas-Fajardo L, Bozzo WR, Bravo ML, Builes JJ, Capilla J, Carvalho M, Castillo C, Catanesi CI, Corach D, Di Lonardo AM, Espinheira R, Fagundes de Carvalho E, Farfán MJ, Figueiredo HP, Gomes I, Lojo MM, Marino M, Pinheiro MF, Pontes ML, Prieto V, Ramos-Luis E, Riancho JA, Souza Góes AC, Santapa OA, Sumita DR, Vallejo G, Vidal Rioja L, Vide MC, Vieira da Silva CI, Whittle MR, Zabala W, Zarrabeitia MT, Alonso A, Carracedo A, Amorim A (2005) Mutation rates at Y chromosome specific microsatellites. Hum Mutat 26: 520-8

Haak W, Balanovsky O, Sánchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, Der Sarkissian CS, Brandt G, Schwarz C, Nicklisch N, Dresely V, Fritsch B, Balanovska E, Villems R, Meller H, Alt KW, Cooper A (2010) Ancient DNA from European early neolithic farmers reveals their near eastern affinities. PLoS Biol 8: e1000536

Haak W, Forster P, Bramanti B, Matsumura S, Brandt G, Tanzer M, Villems R, Renfrew C, Gronenborn D, Alt KW, Burger J (2005) Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. Science 310: 1016-8

Hagelberg E, Goldman N, Lió P, Whelan S, Schiefenhövel W, Clegg JB, Bowden DK (1999) Evidence for mitochondrial DNA recombination in a human population of island Melanesia. Proc Biol Sci 266: 485-92

Hammer MF, Chamberlain VF, Kearney VF, Stover D, Zhang G, Karafet T, Walsh B, Redd AJ (2005) Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. Forensic Sci Int 164: 45-55

Hammer MF, Karafet TM, Park H, Omoto K, Harihara S, Stoneking M, Horai S (2006) Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet 51: 47-58

Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL (2001) Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 18: 1189-203

Harpending HC, Sherry ST, Rogers AR, Stoneking M (1993) The genetic structure of ancient human populations. Curr Anthropol 34: 483-496

Harris DR (1976) Traditional system of plant food production and the origins of agriculture in west Africa. In: Harlan J, De Wet J, Stemler A, editors. Origins of African plant domestication. The Hague: Mouton.: 311-356

Hedman M, Pimenoff V, Lukka M, Sistonen P, Sajantila A (2004) Analysis of 16 Y STR loci in the Finnish population reveals a local reduction in the diversity of male lineages. Forensic Sci Int 142: 37-43

Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 70: 1152-71

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6: 799-803

Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC (1984) DNA sequences from the quagga, an extinct member of the horse family. Nature 312: 282-284

Hitti P (1990) The Arabs: A Short History.

Hofreiter M, Capelli C, Krings M, Waits L, Conard N, Münzel S, Rabeder G, Nagel D, Paunovic M, Jambrĕsić G, Meyer S, Weiss G, Pääbo S (2002) Ancient DNA analyses reveal high mitochondrial DNA sequence diversity and parallel morphological evolution of late pleistocene cave bears. Mol Biol Evol 19: 1244-50

Hofreiter M, Rabeder G, Jaenicke-Després V, Withalm G, Nagel D, Paunovic M, Jambrĕsić G, Pääbo S (2004) Evidence for reproductive isolation between cave bear populations. Curr Biol 14: 40-3

Horai S, Hayasaka K (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am J Hum Genet 46: 828-42

Howell N, Kubacka I, Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? Am J Hum Genet 59: 501-9

Hurles ME, Nicholson J, Bosch E, Renfrew C, Sykes BC, Jobling MA (2002) Y chromosomal evidence for the origins of oceanic-speaking peoples. Genetics 160: 289-303

Ingman M, Gyllensten U (2001) Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. J Hered 92: 454-61

Ingman M, Gyllensten U (2003) Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. Genome Res 13: 1600-6

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

Jakubiczka S, Arnemann J, Cooke HJ, Krawczak M, Schmidtke J (1989) A search for restriction fragment length polymorphism on the human Y chromosome. Hum Genet 84: 86-8

Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific 'fingerprints' of human DNA. Nature 316: 76-9

Jin L, Macaubas C, Hallmayer J, Kimura A, Mignot E (1996) Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. Proc Natl Acad Sci U S A 93: 15285-8

Jobling MA (2001a) In the name of the father: surnames and genetics. Trends Genet 17: 353-7

Jobling MA (2001b) Y-chromosomal SNP haplotype diversity in forensic analysis. Forensic Sci Int 118: 158-62

Jobling MA, Hurles ME, Tyler-Smith C (2004) Human Evolutionary Genetics. Origins, people and disease. Garland Science, New York and Abingdon.

Jobling MA, Pandya A, Tyler-Smith C (1997) The Y chromosome in forensic analysis and paternity testing. Int J Legal Med 110: 118-24

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends Genet 11: 449-56

Jobling MA, Tyler-Smith C (2000) New uses for new haplotypes the human Y chromosome, disease and selection. Trends Genet 16: 356-62

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

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, Redd AJ, Zegura SL, Hammer MF (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet 69: 615-28

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18: 830-8

Kayser M, Brauer S, Weiss G, Schiefenhövel W, Underhill P, Shen P, Oefner P, Tommaseo-Ponzetta M, Stoneking M (2003) Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am J Hum Genet 72: 281-302

Kayser M, Brauer S, Weiss G, Schiefenhövel W, Underhill PA, Stoneking M (2001a) Independent histories of human Y chromosomes from Melanesia and Australia. Am J Hum Genet 68: 173-190

Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M, Honda K, Jobling M, Krawczak M, Leim K, Meuser S, Meyer E, Oesterreich W, Pandya A, Parson W, Penacino G, Pérez-Lezaun A, Piccinini A, Prinz M, Schmitt C, Roewer L, et al. (1997a) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110: 125-33, 141-9

Kayser M, de Knijff P, Dieltjes P, Krawczak M, Nagy M, Zerjal T, Pandya A, Tyler-Smith C, Roewer L (1997b) Applications of microsatellite-based Y chromosome haplotyping. Electrophoresis 18: 1602-7

Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, Mehdi SQ, Rosser Z, Stoneking M, Jobling MA, Sajantila A, Tyler-Smith C (2004) A comprehensive survey of human Y-chromosomal microsatellites. Am J Hum Genet 74: 1183-97

Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V, Gehrig C, Bernini LF, Jespersen J, Bakker E, Roewer L, de Knijff P (2001b) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. Am J Hum Genet 68: 990-1018

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Krüger C, Krawczak M, Nagy M, Dobosz T, Szibor R, de Knijff P, Stoneking M, Sajantila A (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66: 1580-8

Kayser M, Sajantila A (2001) Mutations at Y-STR loci: implications for paternity testing and forensic analysis. Forensic Sci Int 118: 116-21

Ke Y, Su B, Song X, Lu D, Chen L, Li H, Qi C, Marzuki S, Deka R, Underhill P, Xiao C, Shriver M, Lell J, Wallace D, Wells RS, Seielstad M, Oefner P, Zhu D, Jin J, Huang W, Chakraborty R, Chen Z, Jin L (2001) African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. Science 292: 1151-3

Kimura M (1983) The Neutral Theory of Molecular Evolution.

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75: 752-70

Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk HV, Stepanov V, Gölge M, Usanga E, Papiha SS, Cinnioğlu C, King R, Cavalli-Sforza L, Underhill PA, Villems R (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet 72: 313-32

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373-87

Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, Mountain JL (2003) African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr Biol 13: 464-73

Kong QP, Bandelt HJ, Sun C, Yao YG, Salas A, Achilli A, Wang CY, Zhong L, Zhu CL, Wu SF, Torroni A, Zhang YP (2006) Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15: 2076-86

Kraytsberg Y, Schwartz M, Brown TA, Ebralidse K, Kunz WS, Clayton DA, Vissing J, Khrapko K (2004) Recombination of human mitochondrial DNA. Science 304: 981

Lambert DM, Ritchie PA, Millar CD, Holland BJ, Drummond A, Baroni C (2001) Rates of evolution in ancientDNA from Adélie penguins. Science 295: 2270-2273

Lebeuf JP (1981) Du rôle de l'archeology dans le connaissance du Cameroun. In: Tardits C, editor. Contribution de la re´cherche ethnologique a0 l'histoire des civilisations du Cameroun. Paris: CNRS.: 100-125

Leonard JA, Wayne RK, Cooper A (2000) Population genetics of ice age brown bears. Proc Natl Acad Sci U S A 97: 1651-4

Leonard JA, Wayne RK, Wheeler J, Valadez R, Guillén S, Vilà C (2002) Ancient DNA evidence for Old World origin of New World dogs. Science 298: 1613-6

Lessig R, Willuweit S, Krawczak M, Wu FC, Pu CE, Kim W, Henke L, Henke J, Miranda J, Hidding M, Benecke M, Schmitt C, Magno M, Calacal G, Delfin FC, de Ungria MC, Elias S, Augustín C, Tun Z, Honda K, Kayser M, Gusmão L, Amorim A, Alves C, Hou Y, Keyser C, Ludes B, Klintschar M, Immel UD, Reichenpfader B, Zaharova B, Roewer L (2003) Asian online Y-STR Haplotype Reference Database. Leg Med (Tokyo) 5 Suppl 1: S160-3

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

Loreille O, Orlando L, Patou-Mathis M, Philippe M, Taberlet P, Hänni C (2001) Ancient DNA analysis reveals divergence of the cave bear, Ursus spelaeus, and brown bear, Ursus arctos, lineages. Curr Biol 11: 200-3

Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioğlu C, Roseman C, Underhill PA, Cavalli-Sforza LL, Herrera RJ (2004) The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. Am J Hum Genet 74: 532-44

Maca-Meyer N, González AM, Larruga JM, Flores C, Cabrera VM (2001) Major genomic mitochondrial lineages delineate early human expansions. BMC Genet 2: 13

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

Mahtani MM, Willard HF (1993) A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: implications for mechanisms of mutation at short tandem repeat loci. Hum Mol Genet 2: 431-7

Malaspina P, Persichetti F, Novelletto A, Iodice C, Terrenato L, Wolfe J, Ferraro M, Prantera G (1990) The human Y chromosome shows a low level of DNA polymorphism. Ann Hum Genet 54: 297-305

Malyarchuck BA, Rogozin IB (2004) On the Etruscan mitochondrial DNA contribution to modern humans. Am J Hum Genet 75: 920-923; author reply 923-7

Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Drobnič K, Miścicka-Śliwka D (2008) Mitochondrial DNA phylogeny in Eastern and Western Slavs. Mol Biol Evol 25: 1651-8

McIntosh SK, McIntosh RJ (1983) Current direction in West African prehistory. Annu Rev Anthropol 12: 215-258

Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of the human gene frequecies in Europeans. Science 201: 786-792

Messina F, Scorrano G, Labarga CM, Rolfo MF, Rickards O Mitochondrial DNA variation in an isolated area of Central Italy. Ann Hum Biol 37: 385-402

Mills KA, Even D, Murray JC (1992) Tetranucleotide repeat polymorphism at the human alpha fibrinogen locus (FGA). Hum Mol Genet 1: 779

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100: 171-6

Mohammadou E (1973) L'implantation des Peuls au Nord Cameroun. . Paris: CNRS

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297-304

Nasidze I, Quinque D, Dupanloup I, Rychkov S, Naumova O, Zhukova O, Stoneking M (2004) Genetic evidence concerning the origins of South and North Ossetians. Ann Hum Genet 68: 588-99

Norman D (1975) The Arabs and Medieval Europe. London, Uk: Longmann Group Limited.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456: 98-101

Ohta T, Kimura M (1973) The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet Res 22: 201-204

Onofri V, Alessandrini F, Turchi C, Fraternale B, Buscemi L, Pesaresi M, Tagliabracci A (2007) Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis. Int J Legal Med 121: 234-7

Opel KL, Chung DT, Drábek J, Tatarek NE, Jantz LM, McCord BR (2006) The application of miniplex primer sets in the analysis of degraded DNA from human skeletal remains. J Forensic Sci 51: 351-6

Ottoni C, Martinez-Labarga C, Loogvali EL, Pennarun E, Achilli A, De Angelis F, Trucchi E, Contini I, Biondi G, Rickards O (2009a) First genetic insight into Libyan Tuaregs: a maternal perspective. Ann Hum Genet 73: 438-48

Ottoni C, Martinez-Labarga C, Vitelli L, Scano G, Fabrini E, Contini I, Biondi G, Rickards O (2009b) Human mitochondrial DNA variation in Southern Italy. Ann Hum Biol 36: 785-811

Pääbo S (1985) Molecular cloning of Ancient Egyptian mummy DNA. Nature 314: 644-5

Pääbo S (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Natl Acad Sci U S A 86: 1939-43

Pakendorf B, Stoneking M (2005) Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet 6: 165-83

Pala M, Achilli A, Olivieri A, Kashani BH, Perego UA, Sanna D, Metspalu E, Tambets K, Tamm E, Accetturo M, Carossa V, Lancioni H, Panara F, Zimmermann B, Huber G, Al-Zahery N, Brisighelli F, Woodward SR, Francalacci P, Parson W, Salas A, Behar DM, Villems R, Semino O, Bandelt HJ, Torroni A (2009) Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. Am J Hum Genet 84: 814-21

Pallottino M (1981) Genti e culture dell'Italia preromana. Jouvence.: 136

Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, Holland KA, Weedn VW, Gill P, Holland MM (1997) A high observed substitution rate in the human mitochondrial DNA control region. Nat Genet 15: 363-8

Pascali VL, Dobosz M, Brinkmann B (1999) Coordinating Y-chromosomal STR research for the Courts. Int J Legal Med 112: 1

Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Kashani BH, Ritchie KH, Scozzari R, Kong QP, Myres NM, Salas A, Semino O, Bandelt HJ, Woodward SR, Torroni A (2009) Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol 19: 1-8

Petrukhin KE, Speer MC, Cayanis E, Bonaldo MF, Tantravahi U, Soares MB, Fischer SG, Warburton D, Gilliam TC, Ott J (1993) A microsatellite genetic linkage map of human chromosome 13. Genomics 15: 76-85

Phillips C (2005) Using online databases for developing SNP markers of forensic interest. Methods Mol Biol 297: 83-106

Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Álvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet 1: 273-80

Pichler I, Mueller JC, Stefanov SA, De Grandi A, Beu Volpato C, Pinggera GK, Mayr A, Ogriseg M, Ploner F, Meitinger T, Pramstaller PP (2006) Genetic structure in contemporary South Tyrolean isolated populations revealed by analysis of Y-Chromosome, mtDNA, and *Alu* polymorphisms. Hum Biol 78: 441-464

Pilkington MM, Wilder JA, Mendez FL, Cox MP, Woerner A, Angui T, Kingan S, Mobasher Z, Batini C, Destro-Bisol G, Soodyall H, Strassmann BI, Hammer MF (2008) Contrasting signatures of population growth for mitochondrial DNA and Y chromosomes among human populations in Africa. Mol Biol Evol 25: 517-25

Prinz M, Carracedo A, Mayr WR, Morling N, Parsons TJ, Sajantila A, Scheithauer R, Schmitter H, Schneider PM (2007) DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI). Forensic Sci Int Genet 1: 3-12

Pritchard JK, Seielstad MT, Pérez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16: 1791-8

Quintana-Murci L, Krausz C, McElreavey K (2001) The human Y chromosome: function, evolution and disease. Forensic Sci Int 118: 169-81

Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S, Balanovsky O, Kidd KK, Kidd JR, van der Veen L, Hombert JM, Gessain A, Verdu P, Froment A, Bahuchet S, Heyer E, Dausset J, Salas A, Behar DM (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci U S A 105: 1596-601

Quintáns B, Álvarez-Iglesias V, Salas A, Phillips C, Lareu MV, Carracedo A (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140: 251-7

Reed FA, Tishkoff SA (2006) African human diversity, origins and migrations. Curr Opin Genet Dev 16: 597-605

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt H-J (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am. J. Hum. Genet. 67: 1251-76

Ritchie PA, Millar CD, Gibb GC, Baroni C, Lambert DM (2004) Ancient DNA enables timing of the pleistocene origin and holocene expansion of two adelie penguin lineages in antarctica. Mol Biol Evol 21: 240-8

Roewer L, Arnemann J, Spurr NK, Grzeschik KH, Epplen JT (1992) Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet 89: 389-394

Roewer L, Croucher PJ, Willuweit S, Lu TT, Kayser M, Lessig R, de Knijff P, Jobling MA, Tyler-Smith C, Krawczak M (2005) Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Hum Genet 116: 279-91

Roewer L, Kayser M, de Knijff P, Anslinger K, Betz A, Caglià A, Corach D, Füredi S, Henke L, Hidding M, Kärgel HJ, Lessig R, Nagy M, Pascali VL, Parson W, Rolf B, Schmitt C, Szibor R, Teifel-Greding J, Krawczak M (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. Forensic Sci Int 114: 31-43

Roewer L, Krawczak M, Willuweit S, Nagy M, Alves C, Amorim A, Anslinger K, Augustin C, Betz A, Bosch E, Caglià A, Carracedo A, Corach D, Dekairelle AF, Dobosz T, Dupuy BM, Füredi S, Gehrig C, Gusmão L, Henke J, Henke L, Hidding M, Hohoff C, Hoste B, Jobling MA, Kärgel HJ, de Knijff P, Lessig R, Liebeherr E, Lorente M, Martínez-Jarreta B, Nievas P, Nowak M, Parson W, Pascali VL, Penacino G, Ploski R, Rolf B, Sala A, Schmidt U, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Kayser M (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int 118: 106-13

Rolf B, Keil W, Brinkmann B, Roewer L, Fimmers R (2001) Paternity testing using Y-STR haplotypes: assigning a probability for paternity in cases of mutations. Int J Legal Med 115: 12-5

Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barać L, Peričić M, Balanovsky O, Pshenichnov A, Dion D, Grobei M, Zhivotovsky LA, Battaglia V, Achilli A, Al-Zahery N, Parik J, King R, Cinnioğlu C, Khusnutdinova E, Rudan P, Balanovska E, Scheffrahn W, Simonescu M, Brehm A, Goncalves R, Rosa A, Moisan JP, Chaventre A, Ferak V, Füredi S, Oefner PJ, Shen P, Beckman L, Mikerezi I, Terzić R, Primorac D, Cambon-Thomsen A, Krumina A, Torroni A, Underhill PA, Santachiara-Benerecetti AS, Villems R, Semino O (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. Am J Hum Genet 75: 128-37

Rootsi S, Zhivotovsky LA, Baldovič M, Kayser M, Kutuev IA, Khusainova R, Bermisheva MA, Gubina M, Fedorova SA, Ilumäe AM, Khusnutdinova EK, Voevoda MI, Osipova LP, Stoneking M, Lin AA, Ferak V, Parik J, Kivisild T, Underhill PA, Villems R (2007) A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet 15: 204-11

Rosa A, Brehm A, Kivisild T, Metspalu E, Villems R (2004) MtDNA profile of West Africa Guineans: towards a better understanding of the Senegambia region. Ann Hum Genet 68: 340-52

Rose G, Longo T, Maletta R, Passarino G, Bruni AC, De Benedictis G (2008) No evidence of association between frontotemporal dementia and major European mtDNA haplogroups. Eur J Neurol 15: 1006-8

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, Kučinskas 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 (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 67: 1526-43

Rowley-Conwy P (2009) Human prehistory: Hunting for the earliest farmers. Curr Biol 19: R948-9

Ruitberg CM, Reeder DJ, Butler JM (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res 29: 320-2

Salas A, Bandelt HJ, Macaulay V, Richards MB (2007) Phylogeographic investigations: the role of trees in forensic genetics. Forensic Sci Int 168: 1-13

Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt H-J (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem. Biophys. Res. Commun. 335: 891-9

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo A (2002) The making of the African mtDNA landscape. Am J Hum Genet 71: 1082-111

Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74: 454-65

Sánchez JJ, Børsting C, Hallenberg C, Buchard A, Hernandez A, Morling N (2003) Multiplex PCR and minisequencing of SNPs--a model with 35 Y chromosome SNPs. Forensic Sci Int 137: 74-84

Sánchez JJ, Bríon M, Parson W, Blanco-Verea AJ, Børsting C, Lareu M, Niederstätter H, Oberacher H, Morling N, Carracedo A (2004) Duplications of the Y-chromosome specific loci P25 and 92R7 and forensic implications. Forensic Sci Int 140: 241-50

Sánchez JJ, Endicott P (2006) Developing multiplexed SNP assays with special reference to degraded DNA templates. Nat Protoc 1: 1370-8

Santos FR, Tyler-Smith C (1996) Reading the human Y chromosome: the emerging DNA markers and human genetic history. Braz J Genet 19: 665-670

Scheinfeldt L, Friedlaender F, Friedlaender J, Latham K, Koki G, Karafet T, Hammer M, Lorenz J (2006) Unexpected NRY chromosome variation in Northern Island Melanesia. Mol Biol Evol 23: 1628-41

Schlotterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. Nucleic Acids Res 20: 211-5

Schwartz M, Vissing J (2003) [Paternal inheritance of mitochondrial DNA]. Ugeskr Laeger 165: 3627-30

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20: 278-80

Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ, Zhivotovsky LA, King R, Torroni A, Cavalli-Sforza LL, Underhill PA, Santachiara-Benerecetti AS (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet 74: 1023-34

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 (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290: 1155-9

Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA (2002) Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. Am J Hum Genet 70: 265-8

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA (2006) Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet 78: 202-21

Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, Sher AV, Pybus OG, Gilbert MT, Barnes I, Binladen J, Willerslev E, Hansen AJ, Baryshnikov GF, Burns JA, Davydov S, Driver JC, Froese DG, Harington CR, Keddie G, Kosintsev P, Kunz ML, Martin LD, Stephenson RO, Storer J, Tedford R, Zimov S, Cooper A (2004) Rise and fall of the Beringian steppe bison. Science 306: 1561-5

Shen P, Wang F, Underhill PA, Franco C, Yang WH, Roxas A, Sung R, Lin AA, Hyman RW, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (2000) Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci U S A 97: 7354-9

Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, Chakraborty R, Jin L, Su B (2005) Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. Am J Hum Genet 77: 408-19

Shriver MD, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134: 983-93

Sia EA, Kokoska RJ, Dominska M, Greenwell P, Petes TD (1997) Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol Cell Biol 17: 2851-8

Sigurğardóttir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P (2000) The mutation rate in the human mtDNA control region. Am J Hum Genet 66: 1599-609

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66: 262-78

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423: 825-37

Smith KD, Young KE, Talbot CC, Jr., Schmeckpeper BJ (1987) Repeated DNA of the human Y chromosome. Development 101 Suppl: 77-92

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, Torroni A, Richards MB (2010) The archaeogenetics of Europe. Curr Biol 20: R174-83

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

Sobrino B, Brion M, Carracedo A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int 154: 181-94

Spedini G, Destro-Bisol G, Mondovì S, Kaptué L, Taglioli L, Paoli G (1999) The peopling of sub-Saharan Africa: the case study of Cameroon. Am J Phys Anthropol 110: 143-62

Stenico M, Nigro L, Bertorelle G, Calafell F, Capitanio M, Corrain C, Barbujani G (1996) High Mitochondrial Sequence Diversity in Linguistic Isolates of the Alps. Am J Hum Genet 59: 1363-1375

Stephan W (1989) Tandem-repetitive noncoding DNA: forms and forces. Mol Biol Evol 6: 198-212

Stevanovitch A, Gilles A, Bouzaid E, Kefi R, Paris F, Gayraud RP, Spadoni JL, El-Chenawi F, Beraud-Colomb E (2004) Mitochondrial DNA sequence diversity in a sedentary population from Egypt. Ann Hum Genet 68: 23-39

Stoneking M, Sherry ST, Redd AJ, Vigilant L (1992) New approaches to dating suggest a recent age for the human mtDNA ancestor. Philos Trans R Soc Lond B Biol Sci 337: 167-75

Strand M, Prolla TA, Liskay RM, Petes TD (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365: 274-6

Stumpf MP, Goldstein DB (2001) Genealogical and evolutionary inference with the human Y chromosome. Science 291: 1738-42

Su B, Jin L, Underhill P, Martinson J, Saha N, McGarvey ST, Shriver MD, Chu J, Oefner P, Chakraborty R, Deka R (2000) Polynesian origins: insights from the Y chromosome. Proc Natl Acad Sci U S A 97: 8225-8

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L (1999) Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet 65: 1718-24

Tagliabracci C, Turchi C, Buscemi L, Sassaroli C (2001) Polymorphism of the mitochondrial DNA control region in Italians. International Journal of Legal Medicine 114: 224-228

Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res 12: 4127-38

Thomas MG, Barnes I, Weale ME, Jones AL, Forster P, Bradman N, Pramstaller PP (2008) New genetic evidence supports isolation and drift in the Ladin communities of the South Tyrolean Alps but not an ancient origin in the Middle East. Eur J Hum Genet 16: 124-34

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci U S A 97: 7360-5

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, Mountain JL (2007) History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 24: 2180-95

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ (2006) Harvesting the fruit of the human mtDNA tree. Trends Genet 22: 339-45

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML, Wallace DC (1996) Classification of European mtDNAs from an analysis of three European populations. Genetics 144: 1835-50

Torroni A, Miller JA, Moore LG, Zamudio S, Zhuang J, Droma T, Wallace DC (1994) Mitochondrial DNA analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude. Am J Phys Anthropol 93: 189-99

Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, Macaulay V, Scozzari R (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am J Hum Genet 69: 1348-56

Torroni A, Schurr TG, Yang CC, Szathmary EJ, Williams RC, Schanfield MS, Troup GA, Knowler WC, Lawrence DN, Weiss KM, et al. (1992) Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. Genetics 130: 153-62

Turchi C, Buscemi L, Giacchino E, Onofri V, Fendt L, Parson W, Tagliabracci A (2009) Polymorphisms of mtDNA control region in Tunisian and Moroccan populations: an enrichment of forensic mtDNA databases with Northern Africa data. Forensic Sci. Int. Genet 3: 166-72

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7: 996-1005

Underhill PA, Passarino G, Lin AA, Shen P, Mirazón Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet 65: 43-62

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26: 358-61

Verginelli F, Donati F, Coia V, Boschi I, Palmirotta R, Battista P, Mariani Costantini R, Destro-Bisol G (2003) Variation of the hypervariable region-1 of mitochondrial DNA in central-eastern Italy. J Forensic Sci 48: 443-4

Vernesi C, Caramelli D, Dupanloup I, Bertorelle G, Lari M, Cappellini E, Moggi-Cecchi J, Chiarelli B, Castrì L, Casoli A, Mallegni F, Lalueza-Fox C, Barbujani G (2004) The Etruscans: a population-genetic study. Am J Hum Genet 74: 694-704

Vernesi C, Fuselli S, Castri L, Bertorelle G, Barbujani G (2002) Mitochondrial diversity in linguistic isolates of the Alps: a reappraisal. Hum Biol 74: 725-30

Vilà C, Leonard JA, Gotherstrom A, Marklund S, Sandberg K, Liden K, Wayne RK, Ellegren H (2001) Widespread origins of domestic horse lineages. Science 291: 474-7

Vona G, Ghiani ME, Calo CM, Vacca L, Memmi M, Varesi L (2001) Mitochondrial DNA sequence analysis in Sicily. Am J Hum Biol 13: 576-89

Ward RH, Frazier BL, Dew-Jager K, Pääbo S (1991) Extensive mitochondrial diversity within a single Amerindian tribe. Proc Natl Acad Sci U S A 88: 8720-4

Weber JL (1990) Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics 7: 524-30

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2: 1123-8

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, Su B, Pitchappan R, Shanmugalakshmi S, Balakrishnan K, Read M, Pearson NM, Zerjal T, Webster MT, Zholoshvili I, Jamarjashvili E, Gambarov S, Nikbin B, Dostiev A, Aknazarov O, Zalloua P, Tsoy I, Kitaev M, Mirrakhimov M, Chariev A, Bodmer WF (2001) The Eurasian heartland: a continental perspective on Y-chromosome diversity. Proc Natl Acad Sci U S A 98: 10244-9

White PS, Tatum OL, Deaven LL, Longmire JL (1999) New, male-specific microsatellite markers from the human Y chromosome. Genomics 57: 433-7

Wiegand P, Klintschar M (2002) Population genetic data, comparison of the repeat structure and mutation events of two short STRs. Int J Legal Med 116: 258-61

Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, Hammer MF (2004) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet 36: 1122-5

Willerslev E, Cooper A (2005) Ancient DNA. Proc Biol Sci 272: 3-16

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H, Hammer MF (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur J Hum Genet 13: 867-76

Yao YG, Bravi CM, Bandelt HJ (2004) A call for mtDNA data quality control in forensic science. Forensic Sci Int 141: 1-6

Yao YG, Salas A, Bravi CM, Bandelt HJ (2006) A reappraisal of complete mtDNA variation in East Asian families with hearing impairment. Hum Genet 119: 505-15

YCC (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12: 339-48

Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF (2004) High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. Mol Biol Evol 21: 164-75

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhovel W, Fretwell N, Jobling MA, Harihara S, Shimizu K, Semjidmaa D, Sajantila A, Salo P, Crawford MH, Ginter EK, Evgrafov OV, Tyler-Smith C (1997) Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60: 1174-83

Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, Qamar R, Ayub Q, Mohyuddin A, Fu S, Li P, Yuldasheva N, Ruzibakiev R, Xu J, Shu Q, Du R, Yang H, Hurles ME, Robinson E, Gerelsaikhan T, Dashnyam B, Mehdi SQ, Tyler-Smith C (2003) The genetic legacy of the Mongols. Am J Hum Genet 72: 717-21

Zhivotovsky LA, Underhill PA (2005) On the evolutionary mutation rate at Y-chromosome STRs: comments on paper by Di Giacomo et al. (2004). Hum Genet 116: 529-32

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

Online Bibliography

http://www.ornl.gov/sci/techresouces/Human_Genome/home.shtlm

http://www.yhrd.org

http://www.cstl.nist.gov/biotech/strbase

http://www.snp-y.org/

http://www.ethnologue.com/

http://www.nationalgeographic.com/

Additional Bibliography

Álvarez-Iglesias V, Estudio multidisciplinar de la variabilidada del ADN mitocondrial en poblaciones humanas. Tesis doctoral, Universidade de Santiago de Compostela.

Blanco-Verea A, Linajes del cromosoma Y humano: aplicaciones genético-poblacionales y forenses. Tesis doctoral, Universidade de Santiago de Compostela.

Fondevila-Álvarez M, Desarrollo de paneles de SNPs autosómicos y estudio de su aplicación con fines forenses. Tesis doctoral, Universidade de Santiago de Compostela.