UNIVERSITY OF CALABRIA

Ph.D. in Molecular Bio-pathology

(Disciplinary Field BIO/18-Genetics)

Population genetics of the AKR7A2 gene in humans

Candidate

Tamara Benincasa

Supervisors

Prof. G. Rose

Prof. A. Novelletto

Co-ordinator

Prof. G. De Benedictis

2008

CONTENTS

Sommario **I**

Summary **II**

1. INTRODUCTION **1**

1.1 EXAMPLES OF CONSTRUCTION OF NULL HYPOTHESES AND THEIR TESTING **5**

1.1.1 Inter-specific level **6**

1.1.2 Intra-specific level **8**

1.1.3 DESIGNING A POPULATION GENOMICS STUDY **10**

2. FEATURES OF THE Y CHROMOSOME DIVERSITY **11**

3. FEATURES OF THE Y CHROMOSOME DIVERSITY: PAPER 1

3.1 Y-Chromosomal variation in the Czech Republic **12**

4. FEATURES OF THE Y CHROMOSOME DIVERSITY: PAPER 2

4.1 Tracing past human male movements in northern/eastern Africa and western Eurasia : new clues from Y-chromosomal haplogroups E-M78 and J-M12 **22**

5. AIM OF THE WORK **34**

6. IDENTIFYING THE TARGET GENE **35**

6.1 Metabolism and effects of GHB **39**

7. MATERIALS AND METHODS **42**

7.1. Genomic organization of AKR7A2 gene **42**

7.1.1 Subjects **43**

7.1.2 Experimental procedures **44**

7.1.3 Sequencing **45**

7.1.4 Genotyping **46**

7.1.5 Data analysis. **48**

7.1.6 AKR7A2 in the human lineage in comparison to other mammals **49**

7.1.7 Gene-Haplotype diversity **50**

9. RESULTS AND DISCUSSION **52**

9.1 AKR7A2 in the human lineage in comparison to other mammals

9.1.2 The inter-specific analysis reveals the ancestral state **53**

9.1.3 A preliminary exploration of human variation **54**

9.1.4 Apportionment of diversity **55**

9.1.5 Correlates of AKR7A2: climate **57**

9.1.6 Correlates of AKR7A2: genetics **58**

CONCLUSIONS **60**

ACKNOWLEDGMENTS **63**

REFERENCES **64**

SOMMARIO

Il presente lavoro rappresenta un primo passo verso l'analisi comparata dei pattern di diversità osservati in numerose popolazioni umane per il gene *AKR7A2* (che catalizza la riduzione della succinico semialdeide in gamma idrossi butirrato) e per il cromosoma Y, considerato come un paradigma di neutralità. Le prime sezioni del lavoro presentano i dati di variabilità del cromosoma Y sotto forma di pubblicazioni su riviste internazionali, nelle stesse popolazioni analizzate per *AKR7A2*. L'ultima sezione presenta dati originali, non ancora pubblicati, sulla diversità di *AKR7A2* nelle popolazioni appartenenti al pannello HGDP. Gli articoli e i dati originali aggiungono nuovi elementi all'interpretazione della complessa distribuzione della variabilità genetica, su scala continentale e subcontinentale. I lavori sulla variabilità del cromosoma Y mostrano che l'eterogeneità a livello microgeografico è la regola piuttosto che l'eccezione. Tale quota di variabilità può dimostrarsi un'utile risorsa quando si studia la selezione naturale. La parte originale del lavoro consente di identificare *AKR7A2* come un sistema genetico in cui intervengono fattori di complicazione quali la selezione naturale e la ricombinazione. Il presente lavoro, infine, rappresenta un tentativo di valutare l'efficacia di metodi di analisi alternativi rispetto a quelli classici proposti dalla genetica di popolazione, introducendo una prospettiva inter-specifica.

SUMMARY

The present work represents the commencement of a comparison between the patterns of diversity detected in several human populations at *AKR7A2* (a gene of pharmacogenetics relevance known to catalyze the NADPH-dependent reduction of SSA to GHB in human brain) and the Y chromosome, this latter considered as a paradigm of neutrality. The present work is then organized into sections. The first section (Chapters 2-4) summarizes the properties of the male specific portion of the Y chromosome (MSY) which are useful for population genetics studies. Two chapters (3-4) present published works which explore features of Y chromosome diversity in the same populations also analyzed for *AKR7A2*. The last section (Chapter 6) presents original, yet unpublished data on variation of *AKR7A2* as determined in the same populations, which thus represent the first step towards a comparison between the two sets of data. The entire production of articles and original data presented here contributed in unveiling the complex scenario of distribution of genetic variation, both at a continental and sub-continental scale. In fact, our works on Y chromosome variation show that local heterogeneity is the rule rather than the exception. Local variations thus appear to be a useful resource, especially when natural selection is involved. The original part of the work allows to identify *AKR7A2* as a genetic system in which factors of complication intervene, e.g. natural selection and, recombination. Moreover, the second part of the present work represents an attempt to assay the power of methods of analysis different from those of classical population genetics, introducing an inter-specific perspective.

CHAPTER 1

INTRODUCTION

The identification of alleles that have been subjected to positive natural selection during recent human evolution is one of the challenges in human biology. The best examples derive from candidate genes for which there are prior hypotheses of selection. These studies combined the analysis of interspecific divergence, molecular diversity associated with allelic variants and polymorphism in human populations (for a review see Sabeti et al. 2006).

Most models for the dispersal(s) of early humans out of Africa about 100-50 kya assume that only a subset of the African genetic variation left the continent: the effects are still traceable in the extant distribution of genetic variation in human populations (Excoffier 2002; Garrigan and Hammer 2006). Modern humans spreading in Asia, Europe, Oceania and, later, America encountered different climates, pathogens, toxins and sources of food, which generated differential selective pressures requiring local specific adaptive capabilities. More recent marked changes in population size, population density and cultural conditions could have further modified the effects of selective pressures, depicting a complex scenario for the distribution of genetic variation in human populations (Neel 1962). In fact, genetic variation may have been adaptive in certain environments and more neutral in others; this would have balanced and maintained variation across different environments or populations, resulting in, for example, geographic clines in allele frequencies (Cavalli-Sforza et al. 1994; Chikhi et al. 1998). Recent data confirm this expectation for metabolic genes (Hancock et al. 2008).

Until recent times, genetic variation has been considered mostly neutral (Bamshad and Wooding 2003). According to the neutral theory of molecular evolution, the patterns of protein polymorphism seen in nature are more compatible with the hypothesis that most polymorphisms and fixed differences between species are selectively neutral; thus, polymorphisms are

eliminated or fixed in populations as a consequence of the stochastic effects of genetic drift (Kimura 1985).

In contrast with this model, many recent studies have published nucleotide polymorphism data consistent with a role of natural selection in shaping a major quota of genetic variation distribution, in a directional and/or a balanced fashion. A genome-wide survey of >11,000 genes in 39 subjects of Caucasian and African ancestry led Bustamante et al. (2005) to conclude that variation in coding sequences is mainly shaped by weak negative selection. Recently, the Hap-Map phase II data (International Hap-Map Consortium 2007) supported this view, based on the observation of on excess of rare non-synonymous coding variants.

Analyses of specific genes have produced clear cut results. Evans et al. (2004, 2005) revealed signatures of strong positive selection in the lineage leading to humans in the form of an excess of nonsynonymous substitutions in the coding region of Microcephalin, a gene controlling human brain size. Intra-specific human variation confirms that the gene continues to evolve adaptively in human populations, although the expected pattern of phenotypic variation was not observed (Woods RP, Freimer NB, De Young JA et al (2006) Normal variants of Microcephalin and ASPM do not account for brain size variability (Hum Mol Genet 15:2025-2029).

Nucleotide diversity estimates and LD (Linkage Disequilibrium) simulations indicate that several G6PD alleles may have been recently driven to intermediate frequencies by positive selection in the last 10 ky (Sabeti et al. 2002; Tishkoff et al. 2001), i.e. over a period which is associated with the onset of severe malaria in human populations (Livingstone 1971). Thus, G6PD replacement SNPs may have recently reached intermediate frequencies, but have been maintained by balancing selection for resistance to malaria across global groups where malaria is prevalent (Verrelli et al. 2002).

Bersaglieri et al. (2004) demonstrated that strong positive selection occurred in a large region of the genome that includes the Lactase gene, after the separation of European-derived

populations from Asian- and African-derived populations, likely after the colonization of Europe, in response to new dietary habits related to the introduction of dairy farming (Beja-Pereira et al. 2003). Moreover, positive selection of strong intensity is expected to drive alleles towards fixation in a fast and very geography-dependent manner. This has been recently exemplified by Tishkoff et al. (2007) for the parallel increase in frequency of lactase persistence in Europe and Africa.

Also for genome regions long considered as a paradigm of neutrality, such as mtDNA, the hypothesis of a role of natural selection in shaping the overall diversity has been reconsidered (Mishmar et al. 2003; Ruiz-Pesini et al. 2004).

Due to such increasing evidence, discriminating between selective signatures and neutral, historically shaped, variation becomes crucial to make unbiased inferences on both functional variants and past demographic processes.

EXAMPLES OF CONSTRUCTION OF NULL HYPOTHESES AND THEIR TESTING

The correct interpretation of a neutrality test strongly relies on the formulation of an appropriate null hypothesis against which to test the data for evidence of natural selection. Kimura's theory of neutrally evolving mutations is the backbone on which readily testable null hypotheses (also expanding beyond the species level) have been developed.

Another possibility is to compare many non-neutral hypotheses using a likelihood framework, rather than adopting a simple neutral null hypothesis. The problem existing with these methods is that they are computationally intensive. In fact, as evolutionary models become more complex, more parameters are required, and the information in the data can be spread more thinly amongst them. Consequently, also more data are often required to maintain similar levels of certainty when more complex models introduce new parameters.

INTER-SPECIFIC LEVEL

The neutral rate of nucleotide substitution provides a benchmark that can only be exceeded when positive selection also contributes to the substitution process. As a consequence, the ratio between the rate of the non-synonymous DNA substitutions (those involving an amino acid replacement) per site (Ka) to the estimated rate of synonymous changes (Ks) in the same protein (Ka/Ks), can be calculated for inferring the action of positive selection . Ka/Ks>1 ratios have been taken as evidence for directional selection favouring certain amino acid replacements (Yang and Nielsen . 2000; Yang et al. 1997; Yang and Nielsen 2002).

However, this criterion is considered extremely stringent; in fact the most useful applications of this test have been those restricted to specific functional domains of a protein (Hughes et al, 1988; 1989; 1994). More recently, less stringent criteria have been used, with the introduction of a comparative evaluation of Ka/Ks ratio in extended inter-specific comparisons (Dorus et al. 2004).

The comparison of within-species polymorphism and between-species divergence can be extended also over different loci. Under the hypothesis of neutrality, in fact, the Ka/Ks ratio not only should be equal for different species, but also for different genomic regions across species. The Hudson-Kreitman-Aguadè test (Hudson et al. 1987) attempts to control for differences in neutral mutation rates between two loci (or sequences) that might be caused by differences in the level of selective constraint acting in each locus, under the assumption of no recombination within loci and free recombination between loci.

Genetic variants that alter protein function are usually deleterious and are thus less likely to become common or reach fixation (i.e., 100% frequency) than are mutations that have no functional effect on the protein (i.e., silent mutations). Positive selection over a prolonged period, however, can increase the fixation rate of beneficial function-altering mutations, and such changes can be measured by comparison of DNA sequence between species. The increase

can be detected by comparing the rate of non-synonymous (amino acid–altering) changes with the rate of synonymous (silent) or other presumed neutral changes, by comparison with the rate in other lineages, or by comparison with intraspecies diversity (Bustamante et al. 2005). One extreme example of this kind of signature is found in the gene PRM1, which has 13 non-synonymous and 1 synonymous differences between human and chimpanzee (Sabeti et al. 2006). Statistical tests commonly used to detect this signature include the Ka/Ks test, relative rate tests, and the McDonald-Kreitman test (MK). The McDonald-Kreitman (McDonald and Kreitman 1991) test is a powerful test of neutral molecular evolution; furthermore, it can be used to infer the proportion of substitutions driven by positive adaptive evolution (Charlesworth 1993; Akashi 1999; Fay et al. 2001; Smith and Eyre-Walker 2002). Under the MK test, the pattern of evolution within a species is compared to that between species, for two different types of site. Typically the data are divided into synonymous and nonsynonymous sites.

Similar tests can also be applied to other functional sites, such as non-coding regulatory sequences, and their development is an area of active research. This signature can be detected over a large range of evolutionary time scales. Moreover, it focuses on the beneficial alleles themselves, eliminating ambiguity about the target of selection.

Its power is limited, however, because multiple selected changes are required before a gene will stand out against the background neutral rate of change. It is thus typically possible to detect only ongoing or recurrent selection. In practice, when the human genome is surveyed in this manner, few individual genes will give statistically significant signals, after correction for the large number of genes tested. However, the signature can readily be used to detect positive selection across sets of multiple genes. For example, genes involved in gametogenesis clearly stand out as a class having a high proportion of non-synonymous substitutions (Nielsen et al. 2005; Chimpanzee Sequencing and Analysis Consortium 2005).

INTRA-SPECIFIC LEVEL

Major advances in the description of intra-specific variation derive from the coalescent theory (for a review see Rosenberg and Nordborg 2002; Stephens 2001; Wakeley 2003). This leads to a steady state expected distribution of variants in a given sample, which can be summarized into two measures: Theta(pi) or the mean heterozygosity for nucleotide and Theta(S), a measure of the number of alleles expected in the sample. The two have the same expectation and their comparison is the basis of the measure Tajima's D (TD) (Tajima 1989). Deviations from the neutral hypothesis distort this pattern in different ways. Variations in population size and the presence of positive selection both represent deviations and can overlap their effects. For example, positive selection acting on a specific variant which is being driven to fixation causes this variant to increase its frequency to the exclusion of the remaining variants. This will cause a decrease of the overall level of polymorphism [Theta(pi)] more pronounced than the number of polymorphic alleles, i.e.: a negative TD. On the other hand, a recent population expansion causes more newly arisen variation to survive in the population, generating what is known as a star phylogeny of molecular types. In this case, the number of alleles increases more rapidly than the overall level of polymorphisms, also generating a negative TD. Thus, numerically similar values of TD may originate from different processes, with the notable exception that the demographic effects should be detectable on all genes, whereas the selection effect only on the target locus (Luikart et al. 2003).

Derived (that is, not ancestral) alleles arise by new mutation, and they typically have lower allele frequencies than ancestral alleles. When an allele is strongly beneficial and its frequency grows rapidly (selective sweep), however, derived alleles linked to the beneficial allele can hitchhike to high frequency. Because many of these derived alleles will not reach complete fixation (as a result of an incomplete sweep or recombination of the selected allele during the sweep), positive selection creates a signature of a region containing many high-

frequency derived alleles (Bamshad and Wooding 2003). A good example of this kind of signature is the 10-kb region around the Duffy red cell antigen (FY), which has an excess of high-frequency derived alleles in Africans, thought to be the result of selection for resistance to *P. vivax* malaria (Sabeti et al. 2006). The most commonly used test for derived alleles is the Fay and Wu's H test. In practice, the ancestral allele is inferred from the allele present in closely related species, with the assumption that mutation occurred only once at this position and that it occurred after the two species diverged. Determination of the ancestral allele in humans is facilitated by the availability of the chimpanzee genome sequence and by the growing data from additional primate genomes (Chimpanzee Sequencing and Analysis Consortium 2005). The derived-alleles signature differs from the rare-allele signature discussed above in two important ways. First, different demographic effects are potential confounders [for example, population expansion is a major confounder for rare-alleles tests but not for derived alleles tests]. Second, the signature persists for a shorter period because high frequency derived alleles rapidly drift to or near fixation.

When geographically separate populations are subject to distinct environmental or cultural pressures, positive selection may change the frequency of an allele in one population but not in another. Relatively large differences in allele frequencies between populations (at the selected allele itself or in surrounding variation) may therefore signal a locus that has undergone positive selection. For example, the FY*O allele at the Duffy locus is at or near fixation in sub-Saharan Africa but rare in other parts of the world, an extreme case of population differentiation. Similarly, the region around the LCT locus demonstrates large population differentiation between Europeans and non-Europeans, reflecting strong selection for the lactase persistence allele in Europeans. Commonly used statistics for population differentiation include Fst and *p*excess (Sabeti et al. 2006).

DESIGNING A POPULATION GENOMICS STUDY

When scanning large gene sets for evidences of selective signatures, it is important to have a priori hypotheses or strong candidate genes to test. “Fishing expeditions”, in fact, are susceptible to detecting false positives and drawing erroneous conclusions. When a formal population genomic approach is not applicable, it is anyway possible to test for selection by comparing the pattern of variation at a candidate locus with the genome-wide pattern estimated from a set of neutral markers typed in the same individuals or population, or by comparison with available summary statistics estimated from hundreds of coding and non-coding regions (Stephens et al. 2001a). Whatever is the choice, the strength of this approach relies on the capability to define the baseline of neutral variation against which testing for any selective signature.

Among the most promising candidates to test for a signature of local positive selection are genes that encode drug metabolizing enzymes. Many of these genes show marked differences in allele frequencies between populations, and gene variants have been associated with variable responses to foods and drugs (Wilson et al. 2001).

The present work represents the commencement of a comparison between the patterns of diversity detected in several human populations at AKR7A2 (a gene of pharmacogenetics relevance) and the Y chromosome, this latter considered as a paradigm of neutrality. The present work is then organized into 4 sections. The first section (Chapter 2) summarizes the properties of the male specific portion of the Y chromosome (MSY) which are useful for population genetics studies. Two sections (Chapters 3-4) present published works which explore features of Y chromosome diversity in the same populations also analyzed for AKR7A2. The last section (Chapter 6) presents original, yet unpublished data on variation of AKR7A2 as determined in the same populations, which thus represent the first step towards a comparison between the two sets of data.

CHAPTER 2

FEATURES OF THE Y CHROMOSOME DIVERSITY

The male specific portion (MSY) has several useful properties: 1. the availability of numerous SNPs with a robust phylogeny; 2. the availability of numerous microsatellites to further explore diversity within each SNP-defined lineage; 3. the haploid state in males which makes the haplotype reconstruction immediate; 4 a large accumulation of interpopulation variation, due to its reduced effective population size (expected Fst 4 times larger than autosomal loci).

1. The MSY consists of ~60 Mb of DNA, 30 of which represent the euchromatic portion (Skaletsky et al. 2003). This amount of genetic material contains markers that belong to the same classes observed in the autosomes and thus represent a repertoire larger than in mitochondrial DNA (mtDNA). In the MSY, variation in the modules of alphoid DNA, deletions and inversions of large stretches of DNA are observed, as well as variations of smaller magnitude such as Alu insertions, single nucleotide polymorphisms (SNPs) and variation in the 2-5 bp repeats of microsatellites (STRs). All of the above, alone or in combination, have been used for population studies but SNP and STR are by far the most popular.

In the Y literature, SNPs are often referred to as stable binary or biallelic markers as they arise by mutational events that occur with a very low frequency (of the order of 10^{-8}/base pair/generation). The consequence is that the chance of two consecutive events hitting exactly the same nucleotide pair is very low. Considering more than one position on the same DNA molecule, the particular combination of allelic variants (the haplotype) thus represents a record of all mutational events occurred on the lineage leading to that haplotype. Alleles shared by two haplotypes testify of their common ancestry, whereas alleles which differentiate two haplotypes testify that they belong to lineages that diverged some time in the past and, since then, accumulated a different series of mutations. All haplotypes based on SNPs can be then viewed

as the final branches (leaves) of a phylogenetic tree (Karafet et al.2008). The root of the tree is represented by an haplotype (not necessarily found today and thus to be inferred) carrying the ancestral state at all positions found to be variable today. This is also called Most Recent Common Ancestor (MRCA), to signify that all variation existing when the MRCA existed, or earlier, has gone extinct. Each lineage defined by biallelic markers is referred to as a haplogroup, whereas the term haplotype has been restricted to a combination of alleles at STRs (see below).

2. Another important class of markers is represented by Short Tandem Repeats (STRs). These include loci with different length of the basic repeat and extensive searches for developing them as markers have been performed (Kayser et al. 2004 and refs. therein). In any case, the monophyletic origin of STR alleles cannot be assumed, as any allele of a given size can be generated by a number of events from an entire set of parental alleles. Mutation rates at STR loci are orders of magnitude higher than for SNPs, with a relevant heterogeneity among loci. Estimates of mutation rates can be obtained by a variety of direct methods, i.e. comparison of father's vs. son's haplotypes (see the compilation by Gusmao et al. 2005) as well as changes in allele sizes in deep-rooting pedigrees (Heyer et al. 1997; Bianchi et al. 1998; Foster et al. 1998). Evolutionary methods can be also used. Luca et al. (2005) used coalescent reconstructions and obtained locus-specific values comparable to those of the previous methods. Zhivotovsky et al. (2004) obtained an average mutation rate from population rather than family data, considering known foundation events as starting points for the production of the level of diversity observed today.

3. The Y is known as a chromosome that determines, as a whole, the male sex and has no homologue in the individual's chromosome set. However, from the Mendelian point of view, it consists of three distinct portions, known as PAR1 (Pseudo-autosomal), MSY and PAR2 (for a review see Jobling and Tyler-Smith 2003). The MSY is the only portion that is entirely

transmitted from male to male, being free from homologue-homologue recombination (i.e. with the X chromosome). Typically, the subject's haplotype is transmitted unaltered to his male offspring except when a mutation occurs. This haploid state renders markers of the MSY particularly easy to type as only one allele has to be detected and no phase reconstruction is required.

4. The MSY occupies a special place in population genetics theory, as it is a uniparental marker in much the same way as the mtDNA, but with some peculiar features.

The MSY long revealed a low level of diversity per unit of DNA length (Malaspina et al. 1990; Shen et al. 2000), which was compensated only by specific searches in regions particularly prone to mutation (Wilder et al. 2004a). These findings were interpreted as a signature of possible selection on the chromosome, with a consequent young genealogy (Thomson et al. 2000), even younger than mtDNA. However, in recent years, it is becoming increasingly clear that the role of other factors which affect the evolutionary rate have not been taken in due account (Wilder et al. 2004b). In fact, multiple features of human reproductive and migratory behaviour cause populations to depart from panmixia.

The MSY markers showed the highest quotas of population divergence among continents ever recorded for different portions of the genome (Hammer and Zegura 2002; Romualdi et al. 2002). When the role of mutation rate was kept putatively constant by using NRY markers and markers with a similar sequence from the X chromosome, world populations showed a higher degree of structuring for the MSY at the inter- and intra-continental scale (Scozzari et al. 1997; Karafet et al. 1998). Seielstad et al. (1998) compiled data showing a stronger dependance of between-population differentiation on distance for the MSY as compared to both the autosomes and mtDNA, and proposed the idea of a higher proportion of females than males migrating at each generation, increasing the female-mediated gene flow and reducing divergence for all portions of the genome except the MSY (see also Kayser et al. 2001). Others (Malaspina et al.

2000; Karafet et al. 2001) showed that the same regression changes dramatically depending on the geographic area from which populations originated, replicating findings from large datasets of classical genetic markers (Cavalli-Sforza and Feldman 2003). Finally, Wilder et al. (2004a) dismissed the hypothesis that patterns of genetic structure on the continental and global scales are shaped by the higher rate of migration among females than among males.

CHAPTER 3

FEATURES OF THE Y CHROMOSOME DIVERSITY: PAPER 1

Y-Chromosomal variation in the Czech Republic

In order to analyse the contribution of the Czech population to the Y-chromosome diversity landscape of Europe and to reconstruct past demographic events, we typed 257 males from 5 locations for 21 UEPs. Such sampling has allowed to test the hypothesis of the presence of a line of fault in the frequencies of some haplogroups through central Europe, that had repeatedly been hypothesized in the literature (Kayser et al. 2005). We verified that such line exists, but within the Czech Republic it is not as sharp as at the German-Polish border. Thanks to the joined study of SNP and microsatellites, we succeeded in identifying a signal of demographic expansion of the population, that is to be assigned to the bronze age. This represents a relevant correlation with archaeological evidence. 141 carriers of the 3 most common haplogroups were typed for 10 microsatellites and coalescent analyses applied. Sixteen haplogroups characterized by derived alleles were identified, the most common being R1a-SRY10831 and P-DYS257*(xR1a). The pool of haplogroups within I-M170 represented the third most common clade. Overall, the degree of population structure was low. The ages for Hg I-M170, P-DYS257*(xR1a) and R1a-SRY10831 appeared to be comparable and compatible with their presence during or soon after the LGM. A signal of population growth beginning in

the 1st millennium B.C. was detected. Its similarity among the three most common Hgs indicated that growth was characteristic for a gene pool that already contained all of them. The Czech population appears to be influenced to a very moderate extent by genetic inputs from outside Europe in the post-Neolithic and historical times. Population growth post dated the archaeologically documented introduction of Neolithic technology and the estimated central value coincides with a period of repeated changes driven by the development of metal technologies and the associated social and trade organization.

AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 132:132–139 (2007)

Y-Chromosomal Variation in the Czech Republic

F. Luca,[1] F. Di Giacomo,[2] T. Benincasa,[1] L.O. Popa,[3] J. Banyko,[4] A. Kracmarova,[5] P. Malaspina,[2] A. Novelletto,[1,2]* and R. Brdicka[5]

[1]*Department of Cell Biology, University of Calabria, Rende, Italy*
[2]*Department of Biology, University "Tor Vergata", Rome, Italy*
[3]*National Museum of Natural History, Bucharest, Romania*
[4]*University of P. J. Safarik, Kosice, Slovak Republic*
[5]*Institute for Haematology and Blood Transfusion, Prague, Czech Republic*

KEY WORDS Y chromosome; peopling of Europe; genetic dating; microsatellite variation

ABSTRACT To analyze the contribution of the Czech population to the Y-chromosome diversity landscape of Europe and to reconstruct past demographic events, we typed 257 males from five locations for 21 UEPs. Moreover, 141 carriers of the three most common haplogroups were typed for 10 microsatellites and coalescent analyses applied. Sixteen Hg's characterized by derived alleles were identified, the most common being R1a-SRY_{10831} and P-DYS257*(xR1a). The pool of haplogroups within I-M170 represented the third most common clade. Overall, the degree of population structure was low. The ages for Hg I-M170, P-DYS257*(xR1a), and R1a-SRY_{10831} appeared to be comparable and compatible with their presence during or soon after the LGM. A signal of population growth beginning in the first millennium B.C. was detected. Its similarity among the three most common Hg's indicated that growth was characteristic for a gene pool that already contained all of them. The Czech population appears to be influenced, to a very moderate extent, by genetic inputs from outside Europe in the post-Neolithic and historical times. Population growth postdated the archaeologically documented introduction of Neolithic technology and the estimated central value coincides with a period of repeated changes driven by the development of metal technologies and the associated social and trade organization. Am J Phys Anthropol 132:132–139, 2007.

The male-specific portion of the human Y chromosome (MSY) represents, together with mtDNA, an uniparentally inherited polymorphic system. This property is also associated with the ability of the MSY to detect high levels of structure within and between populations. In fact, not only the male-specific portion of the human Y chromosome (MSY) is represented in numbers 1/4 and 1/3 than the autosomes and the X chromosome, respectively, but it is also influenced by the action of additional factors that further reduce its effective population size. These include variance in family size and heritability of reproductive success (Austerlitz and Heyer, 2000; Heyer et al., 2005), both of which could be influenced by social rank and other cultural features in addition to biological determinants (Zerjal et al., 2003). The overall result is a very fast divergence attributable to stochastic factors which, in populations at the sub-continental scale and in the short term, can largely override limited gene flow and the possible action of natural selection.

Major advances in reconstructing the particular realization of the build-up of MSY diversity in the whole world, as well as in more local populations include: a) the ever-increasing detail of MSY phylogeny based on indel and other binary markers with low recurrence of mutation (Y Chromosome Consortium, 2002; Jobling and Tyler-Smith, 2003); b) the estimation of the antiquity of lineages based on either the infinite site mutation model for binary markers with negligible recurrence of mutation (UEP) or the stepwise mutation model for microsatellite variability into the UEP lineages, with a variety of methods (Wilson and Balding, 1998; De Knijff, 2000; Stumpf and Goldstein, 2001; Di Giacomo et al., 2004; Zhivotovsky et al., 2004; Luca et al., 2005); c) the use of coalescent methods to quantify processes of population growth and infer events of subdivision from the observed distribution of molecular types (Weale et al., 2002; Kasperaviciute et al., 2004).

Evidence for a strong structure of the extant populations of Europe for the MSY has been accumulating by analyses of independent sets of binary indel and Single Nucleotide Polymorphisms (SNP) markers (Rosser et al., 2000; Semino et al., 2000). More recent studies, considering both spatial frequency patterns of haplogroups (Hg) with varying degrees of phyletic affinity and their age estimates, have identified clear patterns in the geographic distributions of haplotypes or Hg's as well as sharp genetic boundaries (Cruciani et al., 2004; Rootsi et al., 2004; Semino et al., 2004). They have proposed models for the underlying population movements, with a mainly prehistorical temporal assignment. In addition, analyses focused on specific Hg's or limited geographic regions displayed the outcome of post-Neolithic population processes (Di Giacomo et al., 2004; Pericic et al., 2005). Finally, the analysis of a large database of micro-

Grant sponsor: PRIN-MIUR 2005.

*Correspondence to: Andrea Novelletto, Department of Biology, University of Rome "Tor Vergata", via della Ricerca Scientifica, snc, 00133 Rome, Italy. E-mail: novelletto@bio.uniroma2.it

Received 21 February 2006; accepted 8 August 2006.

DOI 10.1002/ajpa.20500
Published online 31 October 2006 in Wiley InterScience (www.interscience.wiley.com).

Fig. 1. Map of the Czech Republic showing the sampling locations.

satellite data has confirmed the main picture of the MSY geography in Europe, but has also highlighted the signature of more recent population events (Kayser et al., 2005; Roewer et al., 2005).

Here we report on the results of the characterization of MSY diversity in a population of central Europe as obtained with a sampling scheme that detected population substructure in other European countries (Malaspina et al., 2000; Stefan et al., 2001; Brion et al., 2004; Roewer et al., 2005). Indeed, the area here investigated is crucial to understanding the origin of the present-day genetic landscape of the Continent, as it lies on routes that had inevitably to be involved in a wide range of processes, including repopulation events from glacial refugia, population inputs from the Asian-European border, as well as local expansions triggered by major technological advances.

In this paper we specifically tested the hypothesis that MSY diversity in the Czech Republic retains a signature of a sudden population expansion as documented in the local archaeological record.

MATERIALS AND METHODS

Subjects

We studied an overall number of 257 males collected in the five sampling locations shown in Figure 1. These are arranged along a West-to-East transect in the southern part of the Czech Republic. The linear distances between consecutive locations are, from West to East, 65, 65, 65, and 50 km, respectively. Full informed consent was obtained from all participants in this study.

DNA typings

We used 21 UEP markers, i.e. SRY_{10831}, M78, M201, $DYS221_{136}$, M170, M253, P37, M26, M223, M267, M172, M67, M9, LLY22g, Tat, DYS257(p27), M56, M157, M87, the YAP element insertion/deletion polymorphism, and the rearrangement detected by probe p12f2. These allow the recognition of the haplogroups (Hg's) listed in Table 1. The Y Chromosome Consortium (2002) nomenclature for the I clade was revised according to new phylogeny reported by Underhill et al. (2005). Chromosomes that cannot be assigned to any of the above Hg's are classified as Y*(xA,DE,G,I,J,K).

We used the following sequential typing scheme to determine Hg frequencies. YAP (Hammer and Horai, 1995) and DYS257 (Hammer et al., 1998) were typed in all subjects. p12f2 (Rosser et al., 2000) was typed on all YAP(−)/DYS257(G). All subjects producing a positive p12f2 amplification were typed with multiplex 1 (see below). LLY22g, Tat, and M9 (Underhill et al., 1997; Zerjal et al., 1997) were tested sequentially on all subjects carrying ancestral alleles at all markers in multiplex 1. Finally, all unclassified subjects were tested for SRY_{10831} to detect Hg A. Within each Hg the remaining markers were assayed when appropriate: SRY_{10831} (Kwok et al., 1996) within P-DYS257; M267, M172, and M67 (Malaspina et al., 2001; Underhill et al., 2001) within J-p12f2; M253, M223 (multiplex 2), and P37 (YCC, 2002) within I-M170; $DYS221_{136}$ (Hammer et al., 2001) within G-M201; M78 (Underhill et al., 2000) within DE-YAP; M56, M87, and M157 (Underhill et al., 2000) within R1a-SRY_{10831}. These latter were used to search for additional variation within R1a-SRY_{10831}, as originally described in Asia, but they turned out to be monomorphic and thus uninformative in this population. On the whole, in this scheme Hg P-DYS257*(xR1a) is relatively ill-defined. However it has been shown that the bulk of these chromosomes in Europe are also derived at M173, P25, and M269 (Jobling and Tyler-Smith, 2003).

The YAP insertion and p12f2 rearrangement were typed by PCR, followed by direct visualization on agarose gels. M9, LLy22g, Tat, M78, and DYS257 were typed by restriction with HinfI, HindIII, Hsp92II, AciI, and BanI, respectively. M67, M172, DYS221, and M267 were typed by ASO probe hybridization (Di Giacomo et al., 2003, 2004). P37, M56, M87, and M157 were also typed by ASO hybridization, with probes and conditions listed in Table 2.

Multiplex reactions 1 and 2 were developed as fast assays for three and two loci, respectively (Table 3). Each set of loci was multiplexed in a 20 μl reaction with

TABLE 1. Haplogroup absolute and relative frequencies in the five Czech sampling locations

Haplogroup	Location					
	Klatovy	Pisek	J.Hradec	Trebic	Brno	Total
DE-YAP*(xE3b1)						
N	0	0	1	1	0	2
%	0	0	2.0	2.0	0	0.8
E3b1-M78						
N	2	1	4	3	3	13
%	4.2	1.5	8.2	6.1	6.5	5.1
G-M201*(xG2)						
N	1	0	1	0	0	2
%	2.1	0	2.0	0	0	0.8
G2-P15						
N	3	4	2	2	0	11
%	6.3	6.2	4.1	4.1	0	4.3
I-M170* (xI1,I2a,I2b1)						
N	0	3	2	0	0	5
%	0	4.6	4.1	0	0	1.9
I1-M253						
N	2	5	1	1	4	13
%	4.2	7.7	2.0	2.0	8.7	5.1
I2a-P37						
N	7	6	2	2	2	19
%	14.6	9.2	4.1	4.1	4.3	7.4
I2a1-M26						
N	0	0	0	2	1	3
%	0	0	0	4.1	2.2	1.2
I2b1-M223						
N	3	2	2	0	0	7
%	6.3	3.1	4.1	0	0	2.7
J-p12f2*(xJ1,J2)						
N	0	1	0	0	2	3
%	0	1.5	0	0	4.3	1.2
J1-M267						
N	–	–	–	–	–	–
%	–	–	–	–	–	–
J2-M172*(xJ2f)						
N	1	1	1	3	0	6
%	2.1	1.5	2.0	6.1	0	2.3
J2f-M67						
N	0	1	0	1	1	3
%	0	1.5	0	2.0	2.2	1.2
K-M9*(xP,N)						
N	0	1	0	0	0	1
%	0	1.5	0	0	0	0.4
N-LLY22g*(xN3)						
N	–	–	–	–	–	–
%	–	–	–	–	–	–
N3-Tat						
N	0	2	1	1	0	4
%	0	3.1	2.0	2.0	0	1.6
P-DYS257*(xR1a)						
N	11	19	13	16	13	72
%	22.9	29.2	26.5	32.7	28.3	28.0
R1a-SRY$_{10831}$						
N	17	19	16	17	19	88
%	35.4	29.2	32.7	34.7	41.3	34.2
Y*(xA,DE,G,I,J,K)						
N	1	0	3	0	1	5
%	2.1	0	6.1	0	2.2	1.9
Total						
N	48	65	49	49	46	257
%	100.0	100.0	100.0	100.0	100.0	100.0

2.5 mM $MgCl_2$, 0.125 mM dNTPs, 0.375 ng/μl BSA, and 0.5 U Taq Polymerase for 33 cycles (94°C, 20 s; 56°C, 1 min; 72°C, 1 min). One micro liter of each reaction was then reamplified with one of the original primers and two internal fluorescently labeled allele-specific primers (Table 3) in a 10 μl reaction with 2.0 mM $MgCl_2$, 0.20 mM dNTPs, 0.5 ng/μl BSA, and 0.5 U Taq Polymerase for 18 cycles (94°C, 30 s; 42°C, 1 min; 72°C, 30 s). Two micro liters of the product were mixed with 12 μl of formamide, loaded on an ABI 310 automated sequencer and analyzed with the GeneScan software with Tamra as internal standard. Allele states were identified by black (ancestral) or blue (derived) peaks of locus-specific size.

We also typed 141 subjects carrying Hg's I*(xI1,I2a,I2b1), I1-M253, I2a-P37, I2a1-M26, I2b1-M223, P-DYS257*(xR1a), and R1a-SRY$_{10831}$ with the following microsatellite markers (Butler et al., 2002): YCA2A, YCA2B, YCAIIa, YCAIIb, DYS385A and B, DYS388, DYS391, DYS426, DYS439, DYS460, and H4 (individual haplotype data available at www2.bio.uniroma2.it/biologia/laboratori/lab-geneticaumana/geneuma-pubb.htm).

Data analysis

Diversity indexes and AMOVA computations were obtained with the Arlequin 2.000 package (Schneider et al., 2000). To assay the gradient of Hg differentiation, we carried out spatial autocorrelation analysis using the program SAAP (Sokal and Oden, 1978). We performed several runs using different numbers of distance classes to faithfully represent the geographical distances among samples, yet retaining a meaningful number of comparisons in each class.

Dating estimates of Hg antiquity based on microsatellite diversity were obtained with the program BATWING (Wilson and Balding, 1998) under two different conditions: in the first one, all settings for prior distributions were as described (Arredi et al., 2004; Di Giacomo et al., 2004). In the second one, the priors for α (the rate of increase of population size) and β (the time of start of population growth) were relaxed to UNIFORM (0.0, 0.04) and UNIFORM (0.0, 1.0), respectively, in order to explore the signature of population growth which is present in the data. Mutation rates at the 10 STR loci were given GAMMA (2,1000) as prior. We already validated these settings (Luca et al., 2005) by observing that they are able to predict figures for mutation rates obtained in father-son transmissions (Gusmao et al., 2005) and that using lower priors for mutation rates produces convergent results.

For Hg I, information on the phylogenetic relationships between internal haplogroups was used, with the "infsites" option, which permits only STR trees consistent with the Single Nucleotide Polymorphisms (SNP) data. Two I2a1-M26 haplotypes were not considered, as the multirepeat deletion at YCAIIb found on these chromosomes does not conform to the stepwise mutation model assumed in the method. Each BATWING run consisted of 5,000 Markovian steps, after a warmup of 1,000.

A multidimensional representation of population affinities was obtained by Correspondence Analysis (as implemented in SPSS) on the listing of data from Poland, Germany (Kayser et al., 2005), Croatia, Serbia and Bosnia (Marjanovic et al., 2005) and Italy (our unpublished data). This required some approximation in Hg assignment across studies, e.g. lumping of F*(xJ2,K), G*, J1 and the uncharacterized Hg's into a single category.

RESULTS

Haplogroup frequencies are reported in Table 1. The diversity between population samples was very low, with an undetectable overall Fst. R1a-SRY$_{10831}$ is the single

TABLE 2. Probe sequences and hybridization and washing temperatures for four loci assayed by dot-blot hybridization

Locus	PCR primers	Allele	AS0-PROBE	T (°C)
P37	YCC, 2002	ANC.	TTG GTT CAT AGT GTA AA	44
		DER.	TTG GTT CAC AGT GTA AA	44
M56[1]	Paracchini et al., 2002	ANC.	GAT TAC GAA GAA AGG AG	45
		DER.	GAT TAC GAT GAA AGG AG	45
M87[1]	Paracchini et al., 2002	ANC.	GAA TCT TAT ATT TTT GT	40
		DER.	GAA TCT TAC ATT TTT GT	40
M157[1]	Paracchini et al., 2002	ANC.	AAC AAA AAC AAC CAC AAA T	45
		DER.	AAC AAA AAC CAC CAC AAA T	46

[1] Multiplexed in a single reaction.

TABLE 3. Primer sequences and conditions for the allele-specific assays of multiplex PCR reactions 1 and 2

	Reaction 1		Reaction 2		
	Locus	Primer concentration[1] (μM)	Primers	Primer concentration (μM)	Product size (bp)
Multiplex 1					
	M26	0.10	HEX-ATTCAGTGTTCTCTGC	0.15	83
			FAM-ATTCAGTGTTCTCTGT	0.11	
			CCAGTGGTAAAGTTTTATTACAATTT[2]	0.15	
	M170	0.08	HEX-CTTAAAATCATTGTTCA	0.20	96
			FAM-CTTAAAATCATTGTTCC	0.06	
			CCAATTACTTTCAACATTTAAGACC[3]	0.15	
	M201	0.08	HEX-GTACCTATTACGAAAAC	0.15	152
			FAM-GTACCTATTACGAAAAA	0.15	
			TATGCATTTGTTGAGTATATGTC[2]	0.15	
Multiplex 2					
	M223	0.17	HEX-GATAAATTTACTTACAGTC	0.15	160
			FAM-GATAAATTTACTTACAGTT	0.08	
			CCTTTTTGGATCATGGTTCTT[4]	0.15	
	M253	0.10	HEX-ATAGATAGCAAGTTGAC	0.19	135
			FAM-ATAGATAGCAAGTTGAT	0.15	
			CAGCTCCACCTCTATGCAGTTT[5]	0.15	

[1] For both primers; primer sequences as in Underhill et al. (2000).
[2] Same as forward primer of Underhill et al. (2000).
[3] Same as reverse primer of Underhill et al. (2000).
[4] Same as reverse primer of Underhill et al. (2001).
[5] Same as reverse primer of Cinnioglu et al. (2004).

Hg showing the highest frequency in the easternmost sample of Brno. This is in line with the frequency and continental distribution of this Hg reported by Rosser et al. (2000), who showed a sharp decline over the transect here examined. In our data this trend is detected by spatial autocorrelation (Moran I decreasing from 0.026 to −0.967), though it is not statistically significant. Given the linear arrangement of the samples, we also simply regressed Hg frequency on geographic distance, with insignificant results ($r = 0.296$; P = n.s.). Furthermore, no clear trend in the parallel decrease in the frequency of P-DYS257*(xR1a) was detected ($r = 0.791$; P = n.s.). These results point towards a remarkable homogeneity among the five population samples, that can be treated as a single super-sample.

Haplogroup frequencies in the overall sample are in agreement with those reported in the literature for the population of the Czech Republic or neighbouring nations. E3b1-M78 frequency is in the range (2.9%–8%) reported in countries of central Europe (Cruciani et al., 2004; Semino et al., 2004). The frequency of G-M201 (inclusive of G2) agrees with a pattern of leveled frequencies always lower than 4.4% in central Europe (Semino et al., 2000). We show here that ≫80% of the entire haplogroup G-M201 is accounted for by the derived Hg G2-P15. The G2-P15 frequency among Czechs is slightly lower than that found in continental Italy and Greece (6.3 and 6.6%,

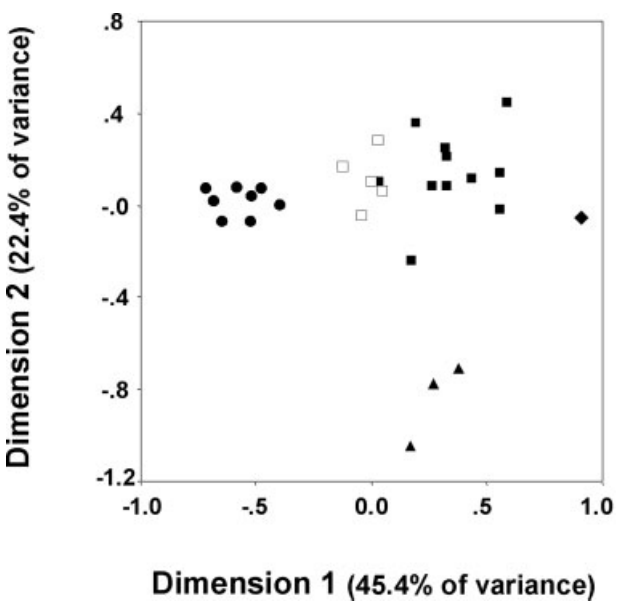


Fig. 2. Correspondence Analysis. Two-dimensional plot of the distribution of populations according to their Hg frequencies. ●: Polish samples; ■: German samples; ◆: Italian sample; ▲: Balkan samples; □: Czech samples.

TABLE 4. STR haplotype diversity for three major Hg's and averages (2.5th to 97.5th percentiles) of posterior distributions from coalescent analyses for Hg age, natural rate of increase of the population (Alpha), and time since beginning of population growth (Beta), under different prior conditions

	I-M170	I1-M253[1]	I2a-P37[1]	I2b1-M223[1]	P-DYS257*(xR1a)	R1a-SRY$_{10831}$
Typed	34	9	14	7	43	62
H[2]	26	7	10	6	36	41
Mean pairwise difference	4.58 ± 2.30	2.03 ± 1.25	1.94 ± 1.17	2.38 ± 1.47	3.52 ± 1.83	2.98 ± 1.58
Diversity	0.978 ± 0.014	0.917 ± 0.092	0.923 ± 0.060	0.952 ± 0.096	0.989 ± 0.008	0.966 ± 0.013
Stringent priors						
Age[3]	534 (261–1,041)	340 (60–1,041)	218 (65–548)	227 (53–623)	581 (22–1,392)	469 (224–1,036)
Alpha	0.040 (0.031–0.050)				0.042 (0.031–0.050)	0.043 (0.031–0.050)
Beta[3]	58 (34–90)				93 (61–141)	84 (53–132)
Relaxed priors						
Age[3]	497 (234–998)	325 (68–998)	218 (72–548)	229 (54–635)	398 (183–889)	348 (176–697)
Alpha	0.023 (0.003–0.039)				0.031 (0.016–0.040)	0.032 (0.015–0.040)
Beta[3]	97 (28–217)				150 (97–222)	125 (82–191)

[1] Age as midpoint between minimum and maximum.
[2] Number of different haplotypes.
[3] In generations.

respectively) (Di Giacomo et al., 2003). The frequencies of Hg I-M170 and its internal Hg's here found are very similar to those previously reported. However, we also found carriers of I2a1-M26, also in this population associated with the diagnostic YCAIIb 11 allele (Ciminelli et al., 1995; Quintana-Murci et al., 1999). This brings the SNP diversity index for the entire I clade to 0.738, i.e. the third highest among the large series of European populations examined (Rootsi et al., 2004). Both Hg's J-p12f2 and N3-Tat display frequencies lower than 5% in the overall sample.

The position of Czech samples in the genetic landscape of central Europe emerges in Correspondence Analysis (Fig. 2). In the first dimension the Czech samples are clustered and positioned in between the Polish and German samples. In fact Hg's R1a and N3-Tat (more frequent in Poland) have the largest negative loadings on this dimension, whereas Hg's K*-M9 and J2*-M172 (more frequent in Germany and Italy) have the largest positive ones. The second dimension is mainly contributed by Hg's I-M170, E3b1-M78, and K*-M9 (negative loadings) vs. Hg P-DYS257*(xR1a). Thus, the position of the three Balkan populations in Figure 2 is in line with the focal distribution of Hg's I and E3b1 in this area (Cruciani et al., 2004; Pericic et al., 2005).

To obtain information on past processes involving the Czech population, we analyzed the diversity at 10 microsatellite loci in the three most common Hg's (Table 4). The 10 loci here used produced a very high STR haplotype diversity, only slightly diminished in I1-M253, I2a-P37, and I2b1-M223. I1-M253 and I2b1-M223 indeed shared YCAII 21-19 as the most common pattern (Rootsi et al., 2004), but this affinity was not extended to the rest of the haplotype, in line with the reviewed phylogeny. In fact, I2b1-M223 shared DYS391, DYS426, and H4 modal alleles with I1-M253, but it also shared DYS388, DYS385b, DYS426, and H4 modal alleles with I2a-P37. Mean pairwise difference among STR haplotypes within each Hg was highest for I-M170, but it was strongly reduced in each of its subclades.

We then applied coalescent dating methods under a model of population growth in the post-glacial period, summarized in the prior distributions for α and β. Under these conditions, the ages of Hg's I-M170, P-DYS257* (xR1a), and R1a-SRY$_{10831}$ turned out to be almost comparable, while the ages of I subclades were remarkably younger than the entire Hg. In our analysis the ages of I2a-P37 and I2b1-M223 were similar and approximately one half that of the entire Hg I-M170 clade, whereas previous results (Rootsi et al., 2004) identified I2b1-M223 (formerly I1c) as the oldest I subclade.

To exclude that strict ranges for α and β priors played an overwhelming role on the reconstruction of the coalescent process, we relaxed the assumptions on these parameters by using flat priors. In these conditions Hg I-M170 as a whole emerges as the oldest, in line with its presence during or soon after the LGM (Rootsi et al., 2004). The results confirmed a growth rate in a narrow range among the three Hg's, between 0.023 and 0.032, i.e. a still marked growth rate though lower than before, paralleled by a more ancient start of growth. The beginning of growth ranges between 97 and 150 generations ago, i.e. between the second and first millennium B.C. by considering 30 years/generation (Arredi et al., 2004). The ages of Hg's P-DYS257*(xR1a) and R1a-SRY$_{10831}$ appear to be only slightly younger but nevertheless place their origins most likely in the Upper Paleolithic. In summary, figures for the rate and start of population growth were similar for the three Hg's, and indicated that growth was characteristic for the whole population in a relatively more recent period.

DISCUSSION

We explored diversity of the male-specific portion of the human Y chromosome (MSY) in five closely spaced Czech population samples. The Hg's P-DYS257*(xR1a) and R1a-SRY$_{10831}$ establish a major divide across central Europe, with a line roughly extending from the Adriatic to the Baltic seas, potentially crossing the transect examined here. This line separates high frequencies of R1a-SRY$_{10831}$ to the East from low frequencies to the West, with an opposite trend for P-DYS257*(xR1a) (Malaspina et al., 2000; Rosser et al., 2000; Semino et al., 2000). The same line was also detected in an analysis of population differentiation at seven microsatellite loci commonly used in forensics (Roewer et al., 2005). Kayser et al. (2005) found this sharp genetic boundary to coincide with the German-Polish border, and interpreted it as the result of massive population movements associated with

World War II, superimposed on pre-existing continent-wide clines. Here we show a very similar composition among five Czech sub-samples. Equating P-DYS257* (xR1a) to R1*(xR1a1), and R1a-SRY_{10831} to R1a1, our sub-samples display frequencies intermediate between the Polish and German ones. The overall ratio between the two Hgs is 1.22 (34.2/28.0) as compared to <1 in Germany and >3 in Poland. Thus, the Czech Republic appears to be affected by a much smoother frequency shift, if any, supporting the interpretation by Kayser et al. (2005) for a very recent origin of the German-Polish discrepancy.

The frequency patterns of other Hg's contribute to the description of the Czech population. First, Hg I-M170 frequency is in the range reported for other central European populations (Semino et al., 2000; Rootsi et al., 2004; Kayser et al., 2005), with the exception of Balkans (Marjanovic et al., 2005). All its three common subclades are represented, and also I2a1-M26 is found, enlarging to the East the area where this latter Hg is found within Europe, though at low frequencies (Capelli et al., 2003; Rootsi et al., 2004). These findings further increase the SNP diversity within Hg I-M170 and raise questions on whether the presence of one or more of its internal haplogroups (before or after the LGM) has been obscured in some areas of Europe by the overlay of other Hg's.

Also, the frequency of YAP+ chromosomes and the large proportion accounted by E3b1-M78 within this clade, is in line with other central European results, signed by Neolithic or post-Neolithic range expansions within the Continent (Cruciani et al., 2004).

The low frequency of J-p12f2 and its subclades confirms the low gene flow with the Mediterranean, from where this Hg spread into southern Europe (Marjanovic et al., 2005; Capelli et al., 2006 and references therein). King and Underhill (2002) associated this Hg with Neolithic inputs into Europe by observing that its presence predicts the distribution of painted pottery. Also, the missing Hg J1-M267 denotes very poor contacts with the Middle East. N3-Tat is a marker of populations who originally spoke Uralic languages, that is found at high frequency in the Baltic region (Zerjal et al., 1997). Its low frequency in the Czech population is in line with a reduced genetic contribution from northern Europe (Jobling and Tyler-Smith, 2003).

Overall, these results identify the Czech population as one influenced to a very moderate extent by genetic inputs from outside Europe in the post-Neolithic and historical times. It thus may represent an ideal population to draw inferences on geographically confined processes that might have occurred also in other parts of central Europe.

Inferences based on STR variation in the three most common Hg's obtained with coalescent methods deserve careful evaluation. First, even though our sampling was carried out in a limited geographic area, it returned age estimates for I-M170, P-DYS257*(xR1a), and R1a-SRY_{10831} similar to those obtained in reports with a wider geographical coverage (Kivisild et al., 2003; Cruciani et al., 2004; Rootsi et al., 2004; Semino et al., 2004; Pericic et al., 2005). These estimates paralleled the trend in mean pairwise difference distribution which, however, weights equally mutation rates across STR loci. Our coalescent estimates are associated with very large confidence intervals (up to three fold the average) and, as such, can only be used with caution to locate in time the age of Hg's (molecules). Conservatively, one can simply conclude that the Czech population harbors a large part of the STR variation generated in each Hg. With two sets of prior assumptions, the ages of the three most common Hg's turned out to be largely overlapping, and compatible with their presence during or soon after the LGM.

However, a local signal emerged from the distribution of this diversity, i.e. that of a fast and recent population growth, which persists even after relaxing the prior assumptions and is similar for the three Hg's. This is summarized by the parameters α and β and their relatively narrow confidence intervals (up to 1.5 fold the average). Estimation of the β parameter most likely locates the beginning of this process in the first millennium B.C., with confidence intervals that are barely compatible with the archaeologically documented introduction of Neolithic technology in this area (Haak et al., 2005). At least for the female lineage, these authors found a little genetic contribution to the present European gene pool from the first farmers settled in the area. Independently from the relevance of these data for reconstructing the genetics of Europe in the early Neolithic (Barbujani and Chicki, 2006), our central value for population growth coincides with a later period of repeated changes in the material cultures in this geographic region, driven by the development of metal technologies and the associated social and trade organization (Childe, 1957; Piggott, 1965; Kristiansen, 1998; Kruta, 2000).

In conclusion, the combined use of SNP and STR markers allowed us to explore different time horizons for the age of molecules and for the process of population growth (Barbujani and Chicki, 2006; Torroni et al., 2006). In fact, our data for the Czech population favor a model in which the age of the most common MSY molecules and their presence in the same gene pool largely predate a consistent population growth. Similar results have been obtained for Lithuania (Kasperaviciute et al., 2004). Both regions lie at the north-western and northern edge, respectively, of the putative homeland (central and south-eastern Europe) of an aboriginal quota of the molecular MSY diversity. This offers an unprecedented opportunity to test alternative models for a continental pattern of diversity which is arranged along the southeast-to-northwest axis (Cavalli-Sforza et al., 1994; Currat and Excoffier, 2005). The question whether this could be the result not only of a single demic diffusion, but also of the demographic increases affecting pre-existing local gene pools is still open (Renfrew, 1979). Examples of recent growth of pre-existing gene pools adding complexity to simple demic diffusion models are provided by mtDNA Hg's HV and H1 (Torroni et al., 1998), as well as Y chromosomal Hg R-SRY_{2627} (Hurles et al., 1999). In order to arrive at a more complete picture, future research efforts should be directed towards the study of additional populations from central and south-eastern Europe with genetic markers and statistical methods which enable inferences on population dynamics. Populations in which a major and recent admixture can be excluded represent an ideal setting to perform this search.

ACKNOWLEDGMENTS

J.B. and L.O.P. were fellows under the NATO "Outreach" programme.

LITERATURE CITED

Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, Makrelouf M, Pascali VL, Novelletto A, Tyler-Smith C. 2004.

A predominantly Neolithic origin for Y-chromosomal DNA variation in North Africa. Am J Hum Genet 75:338–345.

Austerlitz F, Heyer E. 2000. Allelic association is increased by correlation of effective family size. Eur J Hum Genet 8:980–985.

Barbujani G, Chikhi L. 2006. Population genetics: DNAs from the European Neolithic. Heredity 97:84, 85.

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.

Butler JM, Schoske R, Vallone PM, Kline MC, Redd AJ, Hammer MF. 2002. A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. Forensic Sci Int 129: 10–24.

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 13:979–984.

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

Cavalli Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton: Princeton University Press.

Childe VG. 1957. The dawn of European civilization, 6th ed. London: Routledge Kegan Paul.

Ciminelli BM, Pompei F, Malaspina P, Hammer M, Persichetti F, Pignatti PF, Palena A, Anagnou N, Guanti G, Jodice C, Terrenato L, Novelletto A. 1995. Recurrent simple tandem repeat mutations during human Y-chromosome radiation in Caucasian subpopulations. J Mol Evol 41:966–973.

Cinnioglu C, King R, Kivisild T, Kalfoglu 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 114:127–148.

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 74:1014–1022.

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

De Knijff P. 2000. Messages through bottlenecks: On the combined use of slow and fast evolving polymorphic markers on the human Y chromosomes. Am J Hum Genet 67:1055–1061.

Di Giacomo F, Luca F, Anagnou N, Ciavarella G, Corbo RM, Cresta M, Cucci F, Di Stasi L, Agostiano V, Giparaki M, Loutradis A, Mammi' C, Michalodimitrakis EN, 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.

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, Mammi' C, Michalodimitrakis EN, Paoli G, Pappa KI, 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.

Gusmao L, Sanchez-Diz P, Calafell F, Martin P, Alonso CA, Alvarez-Fernandez 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, Farfan MJ, Figueiredo HP, Gomes I, Lojo MM, Marino M, Pinheiro MF, Pontes ML, Prieto V, Ramos-Luis E, Riancho JA, Souza Goes 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 26:520–528.

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 310:1016–1018.

Hammer MF, Horai S. 1995. Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet 56:951–962.

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL. 1998. Out of Africa and back again: Nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15:427–441.

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.

Heyer E, Sibert A, Austerlitz F. 2005. Cultural transmission of fitness: Genes take the fast lane. Trends Genet 21:234–239.

Hurles ME, Veitia R, Arroyo E, Armenteros M, Bertranpetit J, Perez-Lezaun A, Bosch E, Shlumukova M, Cambon-Thomsen A, McElreavey K, Lopez De Munain A, Rohl A, Wilson IJ, Singh L, Pandya A, Santos FR, Tyler-Smith C, Jobling MA. 1999. Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. Am J Hum Genet 65:1437–1448.

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

Kasperaviciute D, Kucinskas V, Stoneking M. 2004. Y chromosome and mitochondrial DNA variation in Lithuanians. Ann Hum Genet 68:438–452.

Kayser M, Lao O, Anslinger K, Augustin C, Bargel G, Edelmann J, Elias S, Heinrich M, Henke J, Henke L, Hohoff C, Illing A, Jonkisz A, Kuzniar P, Lebioda A, Lessig R, Lewicki S, Maciejewska A, Monies DM, Pawlowski R, Poetsch M, Schmid D, Schmidt U, Schneider PM, Stradmann-Bellinghausen B, Szibor R, Wegener R, Wozniak M, Zoledziewska M, Roewer L, Dobosz T, Ploski R. 2005. Significant genetic differentiation between Poland and Germany follows present-day political borders as revealed by Y-chromosome analysis. Hum Genet 117:428–443.

King R, Underhill PA. 2002. Congruent distribution of Neolithic painted pottery and ceramic figurines with Y-chromosome lineages. Antiquity 76:707–714.

Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk HV, Stepanov V, Golge M, Usanga E, Papiha SS, Cinnioglu 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 72:313–332.

Kristiansen K. 1998. Europe before history. Cambridge, UK: Cambridge University Press.

Kruta V. 2000. Les Celtes Histoire et dictionnaire. Paris: Robert Laffont S A.

Kwok C, Tyler-Smith C, Mendonca BB, Hughes I, Berkovitz GD, Goodfellow PN, Hawkins JR. 1996. Mutation analysis of the 2 kb 5′ to SRY in XY females and XY intersex subjects. J Med Genet 33:465–468.

Luca F, Basile M, Di Giacomo F, Novelletto A. 2005. Independent methods for evolutionary genetic dating provide insights into Y-chromosomal STR mutation rates confirming data from direct father-son transmissions. Hum Genet 118:153–165.

Malaspina P, Cruciani F, Santolamazza P, Torroni A, Pangrazio A, Akar N, Bakalli V, Brdicka R, Jaruzelska J, Kozlov A, Malyarchuk B, Mehdi SQ, Michalodimitrakis E, Varesi L, Memmi MM, Vona G, Villems R, Parik J, Romano V, Stefan M, Stenico M, Terrenato L, Novelletto A, Scozzari R. 2000. Patterns of male-specific inter-population divergence in Europe West Asia and North Africa. Ann Hum Genet 64:395–412.

Malaspina P, Tsopanomichalou M, Duman T, Stefan M, Silvestri A, Rinaldi B, Garcia O, Giparaki M, Plata E, Kozlov AI, Barbujani G, Vernesi C, Papola F, Ciavarella G, Kovatchev D, Kerimova MG, Anagnou N, Gavrila L, Veneziano L, Akar N, Loutradis A,

Michalodimitrakis EN, Terrenato L, Novelletto A. 2001. A multistep process for the dispersal of a Y chromosomal lineage in the Mediterranean area. Ann Hum Genet 65:339–349.

Marjanovic D, Fornarino S, Montagna S, Primorac D, Hadziselimovic R, Vidovic S, Pojskic N, Battaglia V, Achilli A, Drobnic K, Andjelinovic S, Torroni A, Santachiara-Benerecetti AS, Semino O. 2005. The peopling of modern Bosnia-Herzegovina: Y-chromosome haplogroups in the three main ethnic groups. Ann Hum Genet 69:757–763.

Paracchini S, Arredi B, Chalk R, Tyler-Smith C. 2002. Hierarchical high throughput SNP genotyping of the human Y chromosome using MALDI-TOF mass spectrometry. Nucleic Acids Res 30:e27.

Pericic M, Lauc LB, Klaric IM, Rootsi S, Janicijevic B, Rudan I, Terzic R, Colak I, Kvesic A, Popovic D, Sijacki A, Behluli I, Dordevic D, Efremovska L, Bajec DD, Stefanovic BD, Villems R, Rudan P. 2005. High-resolution phylogenetic analysis of southeastern Europe traces major episodes of paternal gene flow among Slavic populations. Mol Biol Evol 22:1964–1975.

Piggott S. 1965. Ancient Europe from the beginnings of agriculture to classical antiquity. Edinburgh: Edinburgh University Press.

Quintana-Murci L, Semino O, Poloni ES, Liu A, Van Gijn M, Passarino G, Brega A, Nasidze IS, Maccioni L, Cossu G, al-Zahery N, Kidd JR, Kidd KK, Santachiara-Benerecetti AS. 1999. Y-chromosome specific YCAII, DYS19 and YAP polymorphisms in human populations: A comparative study. Ann Hum Genet 63:153–166.

Renfrew C. 1979. Before civilization the radiocarbon revolution and prehistoric Europe. Cambridge, UK: Cambridge University Press.

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 116:279–291.

Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barac L, Pericic M, Balanovsky O, Pshenichnov A, Dion D, Grobei M, Zhivotovsky LA, Battaglia V, Achilli A, Al-Zahery N, Parik J, King R, Cinnioglu C, Khusnutdinova E, Rudan P, Balanovska E, Scheffrahn W, Simonescu M, Brehm A, Goncalves R, Rosa A, Moisan JP, Chaventre A, Ferak V, Furedi S, Oefner PJ, Shen P, Beckman L, Mikerezi I, Terzic 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 75:128–137.

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, Corte-Real HB, de Knijff P, Decorte R, Dubrova YE, Evgrafov O, Gilissen A, Glisic S, Golge M, Hill EW, Jeziorowska A, Kalaydjieva L, Kayser M, Kivisild T, Kravchenko SA, Krumina A, Kucinskas V, Lavinha J, Livshits LA, Malaspina P, Maria S, McElreavey K, Meitinger TA, Mikelsaar AV, Mitchell RJ, Nafa K, Nicholson J, Norby S, Pandya A, Parik J, Patsalis PC, Pereira L, Peterlin B, Pielberg G, Prata MJ, Previdere 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 within Europe is clinal and influenced primarily by geography rather than language. Am J Hum Genet 66:1526–1543.

Schneider S, Roessli D, Excoffier L. 2000. ARLEQUIN v 2000: A software for population genetics data analysis. Switzerland: Genetics and Biometry Laboratory, University of Geneva.

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 74:1023–1034.

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* in extant Europeans: A Y chromosome perspective. Science 290:1155–1159.

Sokal RR, Oden NL. 1978. Spatial autocorrelation analysis in biology. I Methodology. Biol J Linn Soc 10:199–228.

Stefan M, Stefanescu G, Gavrila L, Terrenato L, Jobling MA, Malaspina P, Novelletto A. 2001. Y chromosome analysis reveals a sharp genetic boundary in the Carpathian region. Eur J Hum Genet 9:27–33.

Stumpf MPH, Goldstein DB. 2001. Genealogical and evolutionary inference with the human Y chromosome. Science 291: 1738–1742.

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

Torroni A, Bandelt HJ, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus ML, Bonne-Tamir B, Scozzari R. 1998. mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62:1137–1152.

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, King RJ, Chow C-ET, Kivisild T, Rootsi S. 2005. New phylogenetic relationships for Y chromosome haplogroup I: Reappraising its pattern of regional subdivision. In: Mellars P, Stringer C, Bar-Yosef O, Boyle K, editors. Rethinking the human revolution: New behavioural and biological perspectives on the origin and dispersal of modern humans. Cambridge, UK: McDonald Institute for Archaeological Research. p 73–75.

Underhill PA, Passarino G, Lin AA, Shen P, Mirazon 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 26:358–361.

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

Wilson IJ Balding DG. 1998. Genealogical inference from microsatellite data. Genetics 150:499–510.

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

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 60:1174–1183.

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 72:717–721.

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

CHAPTER 4

FEATURES OF THE Y CHROMOSOME DIVERSITY: PAPER 2

Tracing past human male movements in northern/eastern Africa and western Eurasia : new clues from Y-chromosomal haplogroups E-M78 and J-M12

Detailed population data were obtained on the distribution of novel biallelic markers that finely dissect the human Y chromosomal haplogroup E-M78. Among 6501 Y chromosomes sampled in 81 human populations worldwide, we found 517 E-M78 chromosomes and assigned them to ten sub-haplogroups. Eleven microsatellite loci were used to further evaluate sub-haplogroup internal diversification. The geographic and quantitative analysis of haplogroup and microsatellite diversity is strongly suggestive of a north-eastern African origin of E-M78, with a corridor for bidirectional migrations between north-eastern and eastern Africa (at least two episodes between 23.9-17.3 ky and 18.0-5.9 ky ago), trans-Mediterranean migrations directly from northern Africa to Europe (mainly in the last 13.0 ky) and flow from north-eastern Africa to western Asia between 20.0 and 6.8 ky ago. A single clade within E-M78 (E-V13) highlights a range expansion in the Bronze Age of south-eastern Europe, which is also detected by haplogroup J-M12. The phylogeography, pattern of molecular radiation and coalescence estimates for both haplogroups are similar and reveal that the genetic landscape of this region is, to a large extent, the consequence of a recent population growth in situ rather than the result of a mere flow of western Asian migrants in the early Neolithic. The results of this paper not only provide a refinement of previous evolutionary hypotheses, but also well defined time frames for past human movements both in northern/eastern Africa and western Eurasia. As far as Europe is concerned this work detected a drastic population burst dated in the third millennium BC in the Southern Balkans, which involved also haplogroup J-M12.

Tracing Past Human Male Movements in Northern/Eastern Africa and Western Eurasia: New Clues from Y-Chromosomal Haplogroups E-M78 and J-M12

Fulvio Cruciani, Roberta La Fratta,* Beniamino Trombetta,* Piero Santolamazza,* Daniele Sellitto,† Eliane Beraud Colomb,‡ Jean-Michel Dugoujon,§ Federica Crivellaro,*[1] Tamara Benincasa,‖ Roberto Pascone,¶ Pedro Moral,# Elizabeth Watson,** Bela Melegh,†† Guido Barbujani,‡‡ Silvia Fuselli,‡‡ Giuseppe Vona,§§ Boris Zagradisnik,‖‖ Guenter Assum,¶¶ Radim Brdicka,## Andrey I. Kozlov,*** Georgi D. Efremov,††† Alfredo Coppa,‡‡‡ Andrea Novelletto,§§§ and Rosaria Scozzari*†*

*Dipartimento di Genetica e Biologia Molecolare, Sapienza Università di Roma, Rome, Italy; †Istituto di Biologia e Patologia Molecolari del Consiglio Nazionale delle Ricerche, Rome, Italy; ‡Laboratoire d'Immunologie, Hôpital de Sainte-Marguerite, Marseille, France; §Laboratoire d'Anthropobiologie, FRE 2960 Centre National de la Recherche Scientifique (CNRS), Université Paul Sabatier, Toulouse, France; ‖Dipartimento di Biologia Cellulare, Università della Calabria, Rende, Italy; ¶Dipartimento di Scienze Ginecologiche Perinatologia e Puericultura, Sapienza Università di Roma, Rome, Italy; #Departament de Biologia Animal, Universitat de Barcelona, Barcelona, Spain; **The Swedish Museum of Natural History, Stockholm, Sweden; ††Department of Medical Genetics and Child Development, University of Pécs, Pécs, Hungary; ‡‡Dipartimento di Biologia, Università di Ferrara, Ferrara, Italy; §§Dipartimento di Biologia Sperimentale, Università di Cagliari, Cagliari, Italy; ‖‖Laboratory of Medical Genetics, General Hospital Maribor, Maribor, Slovenia; ¶¶Institut für Humangenetik, Universität Ulm, Ulm, Germany; ##Institute for Haematology and Blood Transfusion, Prague, Czech Republic; ***ArctAn C Innovative Laboratory, Moscow, Russia; †††Research Center for Genetic Engineering and Biotechnology, Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia; ‡‡‡Dipartimento di Biologia Animale e dell'Uomo, Sapienza Università di Roma, Rome, Italy; and §§§Dipartimento di Biologia, Università ''Tor Vergata'', Rome, Italy

Detailed population data were obtained on the distribution of novel biallelic markers that finely dissect the human Y-chromosome haplogroup E-M78. Among 6,501 Y chromosomes sampled in 81 human populations worldwide, we found 517 E-M78 chromosomes and assigned them to 10 subhaplogroups. Eleven microsatellite loci were used to further evaluate subhaplogroup internal diversification.

The geographic and quantitative analyses of haplogroup and microsatellite diversity is strongly suggestive of a northeastern African origin of E-M78, with a corridor for bidirectional migrations between northeastern and eastern Africa (at least 2 episodes between 23.9–17.3 ky and 18.0–5.9 ky ago), trans-Mediterranean migrations directly from northern Africa to Europe (mainly in the last 13.0 ky), and flow from northeastern Africa to western Asia between 20.0 and 6.8 ky ago.

A single clade within E-M78 (E-V13) highlights a range expansion in the Bronze Age of southeastern Europe, which is also detected by haplogroup J-M12. Phylogeography pattern of molecular radiation and coalescence estimates for both haplogroups are similar and reveal that the genetic landscape of this region is, to a large extent, the consequence of a recent population growth in situ rather than the result of a mere flow of western Asian migrants in the early Neolithic.

Our results not only provide a refinement of previous evolutionary hypotheses but also well-defined time frames for past human movements both in northern/eastern Africa and western Eurasia.

Introduction

A large number of Y chromosome unique event polymorphisms (UEPs) has been reported in the last 7 years (Shen et al. 2000, 2004; Underhill et al. 2000, 2001; Cruciani et al. 2002, 2004, 2006; The Y Chromosome Consortium 2002; Hammer et al. 2003; Cinnioğlu et al. 2004; Rootsi et al. 2004; Semino et al. 2004; Wilder et al. 2004; Kayser et al. 2006; Mohyuddin et al. 2006; Sengupta et al. 2006; Sims et al. 2007) leading to the identification of hundreds of Y-specific haplogroups. Most of the terminal branches of the present Y-phylogenetic tree show a geographic distribution, which is essentially limited to specific continental or subcontinental areas, mainly as a consequence of the reduced effective population size of the Y chromosome and/or of the origin of each branch after major peopling episodes (for a review see Jobling and Tyler-Smith 2003). As previously observed (Cruciani et al. 2004), haplogroup E3b1a (E-M78) escapes this rule, being present at high frequencies in a wide area stretching from northern and eastern Africa, Europe, and western Asia (Underhill et al. 2000, 2001; Bosch et al. 2001, 2006; Cruciani et al. 2002, 2004; Semino et al. 2002, 2004; Arredi et al. 2004; Behar et al. 2004; Cinnioğlu et al. 2004; Flores et al. 2004, 2005; Luis et al. 2004; Shen et al. 2004; Alonso et al. 2005; Gonçalves et al. 2005; Marjanovic et al. 2005; Peričić et al. 2005; Sanchez et al. 2005; Wood et al. 2005; Regueiro et al. 2006).

Due to the lack of informative UEPs defining additional nodes internal to this haplogroup, scholars relied upon the information provided by network analysis of fast evolving microsatellites in order to identify putative monophyletic groups of chromosomes within E-M78 (Cruciani et al. 2004; Semino et al. 2004), an approach which had been successfully used in the past for an initial molecular dissection of major unresolved haplogroups (Malaspina et al. 1998, 2000; Scozzari et al. 1999).

[1] Present address: Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Cambridge, United Kingdom.

Key words: Y-chromosome haplogroups, Y-chromosome phylogeography, human migrations, Bronze Age, European populations, African populations.

E-mail: rosaria.scozzari@uniroma1.it.

Mol. Biol. Evol. 24(6):1300–1311. 2007
doi:10.1093/molbev/msm049
Advance Access publication March 10, 2007

Cruciani et al. (2006) recently reported on the identification of 6 new UEPs within the E-M78 clade, 4 of which seem to be relatively common and informative for evolutionary studies. An evaluation of the correspondence between the subhaplogroups defined by the new UEPs and the E-M78 clusters previously identified by microsatellite network analysis, revealed not only a tight correspondence between the trees generated by the 2 types of markers but also important discrepancies, underlining once more that microsatellite-defined clusters cannot always be considered monophyletic groups of chromosomes (Cruciani et al. 2006).

In the present study, we provide detailed population data on the distribution of E-M78 binary subhaplogroups defined by 10 UEPs (2 of which are here described for the first time) in a sample of 6,501 Y chromosomes belonging to 81 populations mainly from Europe, western Asia, and Africa. In order to obtain estimates of internal diversity and coalescence age of E-M78 subhaplogroups and the associated human migrations and demographic expansions, we also analyzed a set of 11 microsatellites. The same set of microsatellites was also analyzed in a sample of Y chromosomes belonging to the haplogroup J-M12, whose geographic distribution in Europe strictly overlaps that of a single E-M78 subhaplogroup. Our results not only provide a refinement of previous evolutionary hypotheses based on microsatellites alone but also well-defined time frames for different migratory events that led to the dispersal of these haplogroups and subhaplogroups in the Old World.

Subjects and Methods

Subjects

The sample comprised 6,501 unrelated male subjects belonging to 81 populations worldwide. Appropriate informed consent was obtained from all participants. Geographic origin and sample size for each population are reported in table 1 and Supplementary figure 1 (Supplementary Material online).

Molecular Analysis

Samples were obtained from peripheral blood, cultured cells, hair roots, or buccal swabs, and DNA was extracted using appropriate procedures (either phenol–chloroform extraction followed by ethanol precipitation or purification by QlAamp kit from Qiagen, Milan, Italy).

In all, 6,501 Y chromosomes were analyzed for the M78 marker (present study and Cruciani et al. 2002, 2004) by the method of Underhill et al. (2000). Among them, 517 chromosomes carrying the M78-derived T allele were further genotyped for 10 markers defining internal nodes, following a hierarchical approach. Typing methods for 8 of these markers (M148, M224, V12, V13, V19, V22, V27, and V32) were previously described (Underhill et al. 2000, 2001; Cruciani et al. 2006). Two polymorphic markers (V36 and V65) are here reported for the first time. The V36 polymorphism is a T to C transition at position 383 of a 449-bp polymerase chain reaction (PCR) fragment amplified using the primers V36 forward (5′-tcctctttccacttaccttcca) and V36 reverse (5′-caaatgcaaatcaccatttagg). The V65 polymorphism is a G to T transversion at position 77 of a 349-bp PCR fragment amplified using the primers V65 forward (5′-cctcaacctactaaatgtgaccatg) and V65 reverse (5′-atggccacacaattctccat). Both polymorphisms were genotyped by denaturing high performance liquid chromatography. The M12 polymorphism (Underhill et al. 1997), defining haplogroup J2b (Sengupta et al. 2006), has been analyzed as described in Cruciani et al. (2002).

In all, 483 of the 517 E-M78 subjects were further typed for 4 polymorphic dinucleotide repeats (YCAII and DYS413 duplicated loci) and 7 tetranucleotide repeats (DYS19, DYS391, DYS393, DYS439, DYS460 [formerly A7.1], DYS461 [formerly A7.2], and GATA A10) as previously reported (Cruciani et al. 2004). The same eleven microsatellites were analyzed in a set of 43 European J-M12 chromosomes. The DYS392 microsatellite was analyzed in 101 E-M78 chromosomes using primers reported by Butler et al. (2002) and the method described by Cruciani et al. (2002).

Data Analysis

For each haplogroup, phylogenetic relationships among 11 microsatellite haplotypes were obtained by sequentially performing reduced-median and median-joining procedures (Bandelt et al. 1995, 1999) through the use of the network 4.1 program (Fluxus-engineering.com, http://www.fluxus-engineering.com/sharenet.htm). In order to reduce reticulations in the network, microsatellites were weighted proportionally to the inverse of the repeat variance observed in each haplogroup.

To estimate the time to the most recent common ancestor (TMRCA) of haplogroups, we used the 7 tetranucleotide loci and applied the average square distance (ASD) method (Goldstein et al. 1995), where the ancestral haplotype was assumed to be the haplotype carrying the most frequent allele at each microsatellite locus. We employed a microsatellite evolutionarily effective mutation rate (Zhivotovsky et al. 2004). However, because the loci used here and those used by Zhivotovsky et al. (2004) do not overlap completely, we calculated the microsatellite mutation rate as follows: we obtained the mean and standard deviation of the father-to-son mutation rates reported by Gusmão et al. (2005) for the same loci here used, and reduced them by a factor 3.6 (i.e., the discrepancy between the rate estimate obtained from population data and that obtained from father-to-son transmissions [Zhivotovsky et al. 2004]). This resulted in an evolutionarily effective rate $\omega = 7.9 \times 10^{-4}$ (SD $= 5.7 \times 10^{-4}$), a figure that was also used in recalculating the E-M215 coalescence age (data from Cruciani et al. 2004). Recently, Zhivotovsky et al. (2006) showed that reduced loss of diversity in an expanding population brings the evolutionarily effective rate closer to the germ line rate than in constant-size populations. Thus, in the case of expanding populations, we used a correction of the 7.9×10^{-4} value, that was calculated as follows. With reference to figure 2 in Zhivotovsky et al. (2006), the values of accumulated variance in 200–300 generations for the scenarios of 1) a single rate for exponential

Table 1
Frequencies (%) of the Y-Chromosome E-M78 Subhaplogroups in the 81 Populations Analyzed

Population Number	Region and Population	*N*	Frequency of Haplogroup (%)						
			E-M78	E-M78*	E-V12*	E-V13	E-V22	E-V32	E-V65
	Europe								
1	Northern Portuguese[a]	50	4.00	—	—	4.00	—	—	—
2	Southern Portuguese[a]	49	4.08	—	—	4.08	—	—	—
3	Pasiegos from Cantabria[a]	56	—	—	—	—	—	—	—
4	Asturians[a]	90	10.00	—	—	5.56	4.44	—	—
5	Southern Spaniards[a]	62	3.23	—	—	—	3.23	—	—
6	Spanish Basques[a]	55	—	—	—	—	—	—	—
7	French Basques[a,b]	16	6.25	—	6.25	—	—	—	—
8	French[a,b]	225	4.44	—	0.44	4.00	—	—	—
9	English[a,b]	28	—	—	—	—	—	—	—
10	Danish[a]	35	2.86	—	—	2.86	—	—	—
11	Germans	77	3.90	—	—	3.90	—	—	—
12	Polish[a]	40	2.50	—	—	2.50	—	—	—
13	Czechs	268	4.85	—	—	4.85	—	—	—
14	Slovaks	24	8.33	—	—	8.33	—	—	—
15	Slovenians	104	2.88	—	—	2.88	—	—	—
16	Northern Italians[a,b]	94	7.45	—	—	5.32	2.13	—	—
17	Central Italians[a,b]	356	7.87	—	0.28	5.34	1.97	—	0.28
18	Southern Italians[a]	141	10.64	—	0.71	8.51	1.42	—	—
19	Sicilians[c,d]	153	13.07	—	0.65	7.19	4.58	—	0.65
20	Sardinians[a,b,e]	374	3.48	0.27	0.27	1.07	0.80	—	1.07
21	Estonians[a]	74	4.05	—	—	4.05	—	—	—
22	Belarusians	40	—	—	—	—	—	—	—
23	Northern Russians[a,b]	82	3.66	—	—	3.66	—	—	—
24	Southern Russians	92	2.17	—	—	2.17	—	—	—
25	Ukrainians	11	9.09	—	—	9.09	—	—	—
26	Moldovians	77	7.79	—	—	7.79	—	—	—
27	Hungarians	106	9.43	—	—	9.43	—	—	—
28	Rumanians[a]	265	7.55	—	—	7.17	0.38	—	—
29	Macedonians	99	18.18	—	—	17.17	1.01	—	—
30	Continental Greeks	147	19.05	—	—	17.69	0.68	—	0.68
31	Greeks from Crete	215	6.51	—	0.93	5.58	—	—	—
32	Greeks from Aegean Islands	71	16.90	—	—	15.49	1.41	—	—
33	Bulgarians[a]	204	16.67	—	0.49	16.18	—	—	—
34	Albanians[a]	96	32.29	—	—	32.29	—	—	—
	Northwestern Africa								
35	Moroccan Arabs[a]	55	40.00	3.64	—	—	7.27	—	29.09
36	Asni Berbers	54	3.70	—	—	—	3.70	—	—
37	Bouhria Berbers	67	1.49	—	—	1.49	—	—	—
38	Moyen Atlas Berbers[a]	69	10.14	—	—	—	—	—	10.14
39	Marrakech Berbers[a]	29	6.90	—	3.45	—	3.45	—	—
40	Moroccan Jews	50	12.00	—	2.00	2.00	8.00	—	—
41	Mozabite Berbers[a,b]	20	—	—	—	—	—	—	—
	Northeastern Africa								
42	Libyan Jews	25	8.00	—	—	4.00	—	—	4.00
43	Libyan Arabs	10	20.00	—	—	—	—	—	20.00
44	Northern Egyptians (Delta)[a]	72	23.61	—	5.56	1.39	13.89	2.78	—
45	Egyptian Berbers	93	6.45	—	2.15	—	—	—	4.30
46	Egyptians from Baharia	41	41.46	—	14.63	2.44	21.95	—	2.44
47	Egyptians from Gurna Oasis	34	17.65	5.88	8.82	—	—	2.94	—
48	Southern Egyptians[a]	79	50.63	—	44.30	1.27	3.80	—	1.27
	Eastern Africa								
49	Amhara[a]	34	8.82	—	—	—	—	8.82	—
50	Ethiopian Jews[a]	22	9.09	—	—	—	—	9.09	—
51	Mixed Ethiopians[a]	12	33.33	—	—	—	25.00	8.33	—
52	Borana/Oromo (Kenya/Ethiopia)[a]	32	40.63	—	—	—	—	40.63	—
53	Wolayta[a]	12	16.67	—	—	—	8.33	8.33	—
54	Somali[a]	23	52.17	—	—	—	4.35	47.83	—
55	Nilotic from Kenya[a]	18	11.11	—	—	—	11.11	—	—
56	Bantu from Kenya[a,b]	28	3.57	—	—	—	—	3.57	—
57	Western Africa[a,b,f]	123	0.81	—	0.81	—	—	—	—
58	Central Africa[a,b]	150	0.67	—	0.67	—	—	—	—
59	Southern Africa[a,b]	105	—	—	—	—	—	—	—
	Western Asia								
60	Istanbul Turkish[a]	35	8.57	—	—	2.86	5.71	—	—
61	Southwestern Turkish[a]	40	2.50	—	—	2.50	—	—	—

Table 1
Continued

Population Number	Region and Population	*N*	Frequency of Haplogroup (%)						
			E-M78	E-M78*	E-V12*	E-V13	E-V22	E-V32	E-V65
62	Northeastern Turkish[a]	41	—	—	—	—	—	—	—
63	Southeastern Turkish[a]	24	4.17	—	—	4.17	—	—	—
64	Erzurum Turkish[a]	25	4.00	—	4.00	—	—	—	—
65	Central Anatolian[a]	61	6.56	—	1.64	4.92	—	—	—
66	Turkish Cypriots[a]	46	13.04	—	—	10.87	2.17	—	—
67	Sephardi Turkish[a]	19	—	—	—	—	—	—	—
68	Palestinians[a]	29	10.34	—	—	3.45	6.90	—	—
69	Druze Arabs[a]	28	10.71	—	—	10.71	—	—	—
70	Bedouin[a]	28	3.57	—	—	—	3.57	—	—
71	Syrians	100	2.00	—	—	—	2.00	—	—
72	Kurds from Iraq	20	—	—	—	—	—	—	—
73	Arabs from United Arab Emirates[a]	40	2.50	—	—	—	2.50	—	—
74	Omanite[a]	106	0.94	—	—	—	0.94	—	—
75	Adygei[a,b]	18	—	—	—	—	—	—	—
76	Azeri[a]	97	2.06	—	—	2.06	—	—	—
77	Southern Asia[a,b]	300	1.00	—	—	—	1.00	—	—
78	China[a,b]	206	—	—	—	—	—	—	—
79	Eastern Asia[a,b]	41	—	—	—	—	—	—	—
80	Oceania[a,b]	21	—	—	—	—	—	—	—
81	Central and Southern America Native American[a,b]	43	—	—	—	—	—	—	—
	Total	6,501	7.95	0.08	1.00	4.45	1.29	0.54	0.60

[a] This sample, or a subset of it, was previously typed for the M78 marker (Cruciani et al. 2004).
[b] Sample (or a subset of it) from the Human Genome Diversity Project/CEPH DNA panel (Cann et al. 2002).
[c] 43 subjects from Sicily (Trapani) analyzed by Cruciani et al. (2004) are included on the sample.
[d] One E-V13 subject also carries the V27 mutation.
[e] Two E-V22 subjects also carry the V19 mutation.
[f] 106 subjects from Burkina Faso analyzed by Cruciani et al. (2002) are included on the sample.

population growth and 2) growth with 4 distinct consecutive rates were compared with the amount accumulated in constant-size populations. This resulted in evolutionarily effective mutation rates decreased of factors 2.4 and 2.8, respectively (instead of 3.6), that is, 11.9×10^{-4} (SD = 8.5×10^{-4}) and 10.2×10^{-4} (SD = 7.3×10^{-4}), which were applied to haplogroups E-V13 and J-M12 found in Europe. Confidence intervals (CIs) for the ASD (and TMRCA) were obtained as follows: Mutations on the microsatellite genealogy were simulated using a Poisson process, in which the total number of mutational events was calculated based on branch length and assuming that mutations at each microsatellite were gamma-distributed with mean and standard deviation calculated as above. Each mutation increased or decreased allele length by one step (each with probability 0.5). ASD was then evaluated for the simulated data and the whole process repeated 1,000 times to quote the central 95% of values. This method represents a refinement of that by Thomas et al. (1998) and Scozzari et al. (2001), as it also takes into account heterogeneity of mutation rates across loci. An independent dating method (ρ statistics; Forster et al. 1996; Saillard et al. 2000) was also used to assay how robust the time obtained is to choice of method.

Both dating procedures rely on the appropriate choice of a haplotype to be considered ancestral, which remains an uncontrolled source of uncertainty. We observe that the ρ-based ages are slightly younger than the ASD-based ones (fig. 1). The difference is significant only for the root of the entire haplogroup, this being attributable to the relevant departure from a star-like structure because of repeated founder effects (Saillard et al. 2000). Only values obtained from ASD are quoted in the text.

Haplogroup diversity and its sampling variance were estimated as in Arlequin 3.0 (Excoffier et al. 2005).

Frequency and variance maps were depicted on a grid of 44×60 lines using the Kriging procedure (Cressie 1991) through the use of the program Surfer 6.0 (Golden Software, Inc., Golden, CO). The map of microsatellite variances was obtained after pooling data from locations with less than 3 observations and assigning the resulting figures to the centroid of the pooled locations. These points are plotted in figure 5.

Results and Discussion

Molecular Dissection of E-M78 Haplogroup

By analyzing a worldwide sample of 6,501 male subjects, we have identified 517 chromosomes belonging to haplogroup E-M78, more than twice the number found in a previous study (Cruciani et al. 2004). These chromosomes have been further analyzed for the biallelic markers M148 (Underhill et al. 2000), M224 (Underhill et al. 2001), V12, V13, V19, V22, V27, V32 (Cruciani et al. 2006), V36, and V65 (present study). Only 2 of the markers analyzed (V13 and V36) were phylogenetically equivalent, leading to the identification of a total of 10 distinct haplogroups/paragroups (fig. 1), with only 5 chromosomes remaining in the paragroup E-M78*. Four subhaplogroups were either rare (1 and 2 subjects for E-V27 and E-V19, respectively) or absent (E-M148 and E-M224) in the global sample, whereas the

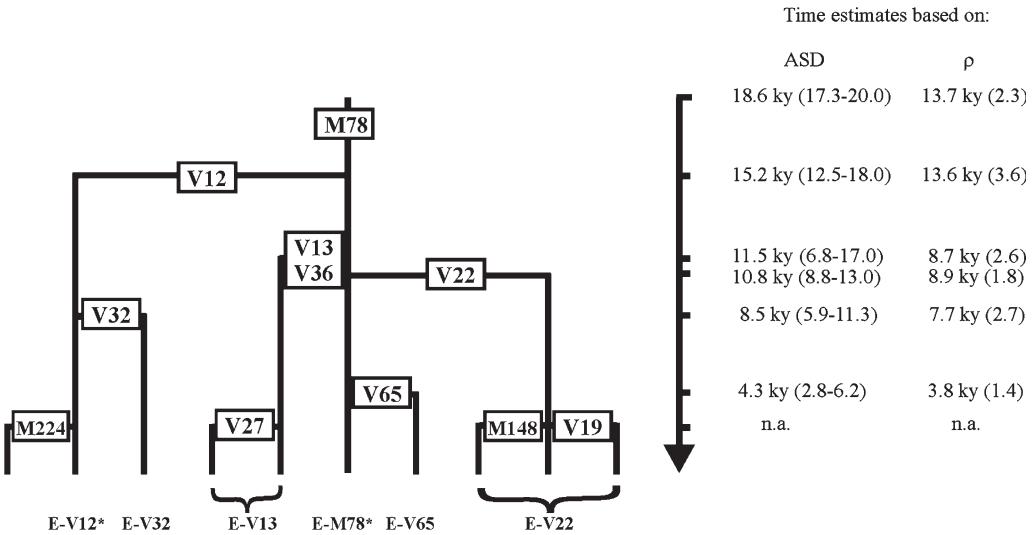


Fig. 1.—Maximum parsimony phylogeny of haplogroup E-M78. Coalescent estimates for the haplogroup E-M78 and major subhaplogroups are shown on the right. The ASD-based estimates are reported with their 95% CIs (in parentheses); the ρ-based estimates are reported with their SD (in parentheses). For haplogroup E-V13, the value obtained on the subset of western Asian samples is reported as this is free from the effect of the population expansion recorded in Europe (see text). Haplogroup nomenclature as cited in the text is reported at the bottom; n.a. —not available (rare haplogroups).

other haplogroups/paragroups were relatively common (table 1 and fig. 2). The E-M78 subhaplogroup identified by the new mutation V65 includes all but 2 of the chromosomes previously included in the cluster β and 1 chromosome from cluster γ of the E-M78 microsatellite network (Cruciani et al. 2004), once again underlining the strong but not perfect correspondence between microsatellite-defined clusters and UEP-defined haplogroups (Cruciani et al. 2006).

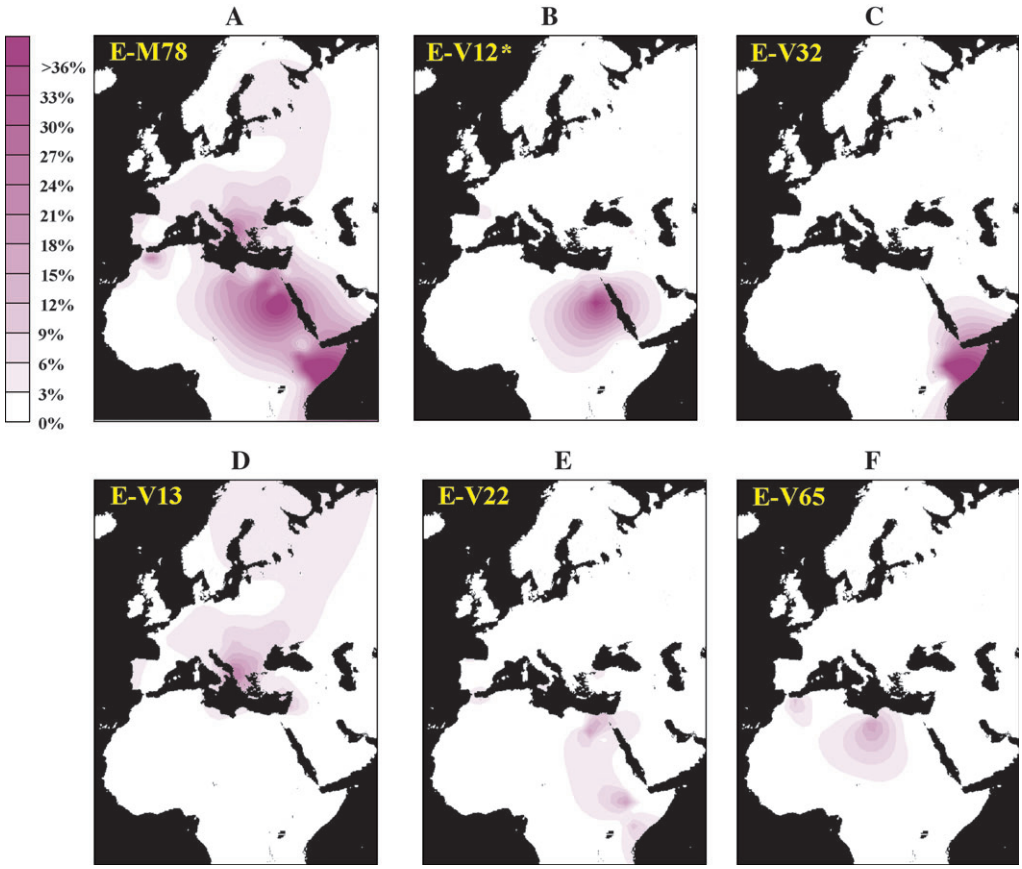


Fig. 2.—Maps of the observed haplogroup/paragroup frequencies. (*A*) E-M78; (*B*) E-V12*; (*C*) E-V32; (*D*) E-V13; (*E*) E-V22; and (*F*) E-V65.

The subdivision of E-M78 in the six common major clades revealed a pronounced geographic structuring (table 1 and fig. 2): Haplogroup E-V65 and the paragroups E-M78* and E-V12* were observed mainly in northern Africa, haplogroup E-V13 was found at high frequencies in Europe, and haplogroup E-V32 was observed at high frequencies only in eastern Africa. The only haplogroup showing a wide geographic distribution was E-V22, relatively common not only in northeastern and eastern Africa but also found in Europe and western Asia, up to southern Asia (table 1, fig. 2).

Locating the Origin of Haplogroup E-M78

An eastern African origin for this haplogroup was hypothesized on the basis of the exclusive presence in that area of a putative ancestral 12-repeat allele at the DYS392 microsatellite, found in association with E-M78 chromosomes (Semino et al. 2004). Northeastern African populations were not represented in that study. In order to test this hypothesis, we analyzed for DYS392, a geographically widespread subset of the E-M78 chromosomes here identified. We observed that the DYS392 12-repeat allele is associated with the majority of the chromosomes belonging to the northeastern African E-V12* (15 out of 18) and to the eastern African E-V32 (21 out of 23), with about half (9 out of 21) of the E-V22 chromosomes (both in eastern and northeastern Africa), with a few of the European E-V13 (2 out of 23), and with some north-African E-V65 (3 out of 16) chromosomes. These findings show that the DYS392 12-repeat allele is common in different regions characterized by high frequencies of E-M78 and suggest that it was most likely generated by multiple mutational events occurring in different UEP-defined subhaplogroups. Thus, the DYS392 allele distribution is not informative to infer the place of origin of E-M78 chromosomes.

An eastern African origin for haplogroup E-M78 was also hypothesized on the basis of the frequency distribution and microsatellite diversity (Cruciani et al. 2004). We may now test this hypothesis by exploiting the new information provided by internal biallelic markers and the extensive resampling in which northeastern Africa is covered by a robust group of 90 E-M78 chromosomes. The frequencies of E-M78 in northeastern Africa and eastern Africa are not significantly different (0.25 ± 0.03 and 0.22 ± 0.02, respectively). As far as the microsatellite diversity is concerned, the highest mean variances across 7 tetranucleotide loci are those observed in eastern Africa and northeastern Africa (0.50 and 0.46, respectively), but an examination of the variances at individual loci reveals that in eastern Africa there is a disproportionate contribution of DYS19 to the mean variance (1.87). This is likely due to a multirepeat deletion associated with the common eastern African E-V32 haplogroup (Cruciani et al. 2006 and supplementary table 1). When this locus is removed from the analysis, we obtain mean variances across 6 loci of 0.41 and 0.27 for northeastern and eastern Africa, respectively. Variances at the 6 individual loci are always higher in the former region, and this difference is statistically significant for the microsatellite locus DYS461 (F test for equality of variances $P < 0.05$). Finally, a greater diversity of E-M78 binary subhaplogroups can be observed in northeastern Africa (0.61 ± 0.04), where all the E-M78 major branches are present, than in eastern Africa (0.30 ± 0.08), where only subhaplogroups E-V22 and E-V32 are found. E-V22 is observed at high frequencies in both northeastern and eastern Africa, with microsatellite variances of 0.46 and 0.35, respectively. The other common eastern African subhaplogroup, E-V32, that represents about 82% of the eastern African E-M78 chromosomes, is a relatively recent terminal branch of E-V12 (8.5 ky, fig. 1), the remaining E-V12 chromosomes being found almost exclusively in northeastern Africa as paragroup E-V12*. The haplogroups E-V13 and E-V65 are also found in northeastern Africa. Although an origin for E-V13 outside the region is likely (see below), E-V65 probably originated in situ as inferred on the basis of its nearly exclusive presence and diversity. It is also worth noting that the rare paragroup E-M78* has not been observed in eastern Africa; moreover, the 2 northwestern African E-M78* chromosomes are well differentiated from the 2 northeastern African E-M78* chromosomes (supplementary table 1, Supplementary Material online) adding a new argument for a higher haplogroup diversity in northern Africa.

In conclusion, the peripheral geographic distribution of the most derived subhaplogroups with respect to northeastern Africa, as well as the results of quantitative analysis of UEP and microsatellite diversity are strongly suggestive of a northeastern rather than an eastern African origin of E-M78. Northeastern Africa thus seems to be the place from where E-M78 chromosomes started to disperse to other African regions and outside Africa.

A Corridor for Bidirectional Migrations between Northeastern and Eastern Africa

The evolutionary processes that determined the wide dispersal of the E-M78 lineages from northeastern Africa to other regions can now be addressed.

E-M78 belongs to clade E3b (E-M215). On the basis of robust phylogeographic considerations, an eastern African origin has been proposed for E-M215 (Underhill et al. 2001; Cruciani et al. 2004), with a coalescence time of 22.4 ky (95% CI 20.9–23.9 ky; recalculated from Cruciani et al. [2004], see Subjects and Methods). A northeastern African origin for haplogroup E-M78 implies that E-M215 chromosomes were introduced in northeastern Africa from eastern Africa in the Upper Paleolithic, between 23.9 ky ago (the upper bound for E-M215 TMRCA in eastern Africa) and 17.3 ky ago (the lower bound for E-M78 TMRCA here estimated, fig. 1). In turn, the presence of E-M78 chromosomes in eastern Africa can be only explained through a back migration of chromosomes that had acquired the M78 mutation in northeastern Africa. The nested arrangement of haplogroups E-V12 and E-V32 defines an upper and lower bound for this episode, that is, 18.0 ky and 5.9 ky, respectively. These were probably not massive migrations, because the present high frequencies of E-V12 chromosomes in eastern Africa are entirely accounted for by E-V32, which most likely underwent subsequent

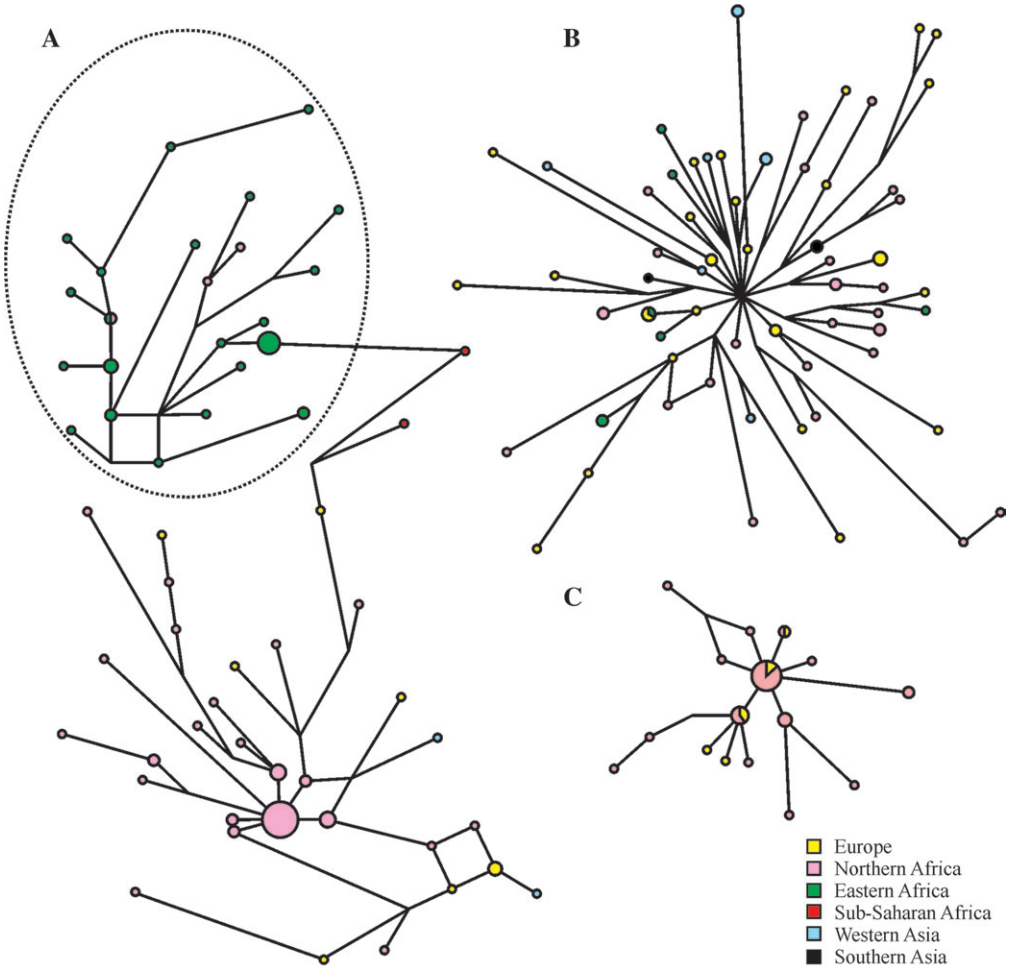


Fig. 3.—Microsatellite networks of haplogroups E-V12 (*A*); E-V22 (*B*); and E-V65 (*C*). In network (*A*), a dotted circle includes all of the E-V12 chromosomes carrying the V32 mutation. Branch lengths are proportional to the number of one-repeat mutations separating 2 haplotypes. Each circle area is proportional to the frequency of the sampled haplotype.

geographically restricted demographic expansions involving well differentiated molecular types (fig. 3*A*). Conversely, the absence of E-V12* chromosomes in eastern Africa is compatible with loss by drift. Possible more recent episodes of gene flow are associated with the less common E-V22 subhaplogroup, also present in both northeastern and eastern Africa, but without a clear pattern of microsatellite molecular differentiation (fig. 3*B*). It is conceivable that the Nile river valley has acted as a genetic corridor for human migrations between northeastern and eastern Africa, a scenario that is also supported by mtDNA analysis both at HV1 (Krings et al. 1999) and entire molecule sequence (Olivieri et al. 2006). There are also other Y-chromosome haplogroups shared by northeastern and eastern African populations, that is, E-M123, J-M267, and K-M70 (Underhill et al. 2000, Cruciani et al. 2004; Luis et al. 2004; Semino et al. 2004; Sanchez et al. 2005). However, unlike E-V12 and E-V22, these haplogroups are also common in western Asia, where they probably originated (Cruciani et al. 2004; Luis et al. 2004; Semino et al. 2004). Thus, it is unclear whether their present geographic distribution in Africa is the consequence of the same evolutionary events that involved the E-M78 chromosomes or whether they have been introduced independently from western Asia in eastern and northeastern Africa. Only the molecular dissection of haplogroups E-M123, J-M267, and K-M70 along with an extensive sampling of populations from these regions will help in answering this question.

Direct Northern African Contribution to the European Gene Pool

Previous studies on the Y-chromosome phylogeography have revealed that central and western Asia were the main sources of Paleolithic and Neolithic migrations contributing to the peopling of Europe (Underhill et al. 2000; Wells et al. 2001). Only sporadic traces of northern African Y chromosomes were found in the European gene pool, mainly linked to the presence at low frequencies of the E-M81 haplogroup in Mediterranean coastal populations (Bosch et al. 2001; Scozzari et al. 2001; Cruciani et al. 2004; Gonçalves et al. 2005). The molecular dissection of E-M78 contributes to the understanding of the genetic relationships between northern Africa and Europe. Several lines of evidence suggest that E-M78 subhaplogroups E-V12, E-V22, and E-V65 have been involved in trans-Mediterranean

migrations directly from Africa. These haplogroups are common in northern Africa, where they likely originated, and are observed almost exclusively in Mediterranean Europe, as opposed to central and eastern Europe (table 1 and fig. 2). Also, among the Mediterranean populations, they are more common in Iberia and south-central Europe than in the Balkans, the natural entry-point for chromosomes coming from the Levant. Such findings are hardly compatible with a southeastern entry of E-V12, E-V22, and E-V65 haplogroups into Europe. Upper limits for the introduction of each of these haplogroups in Europe are given by their estimated ages (fig. 1), whereas lower bounds should be close to the present times, given the lack of internal geographic structuring (fig. 3*A–C*; Cruciani et al. 2004; Semino et al. 2004). Considering both these E-M78 subhaplogroups (present study) and the E-M81 haplogroup (Cruciani et al. 2004), the contribution of northern African lineages to the entire male gene pool of Iberia (barring Pasiegos), continental Italy, and Sicily can be estimated as 5.6%, 3.6%, and 6.6%, respectively. Whether lineages E-M123, J-M267, G-M201, and K-M70, commonly found in both northern Africa and Europe (Bosch et al. 2001; Arredi et al. 2004; Cruciani et al. 2004; Semino et al. 2004), were involved in the same population movements remains to be ascertained due to the poor phylogeographic resolution of these haplogroups.

The Haplogroup E-V13: Migrations and Demographic Expansions in Western Eurasia

Haplogroup E-V13 is the only E-M78 lineage that reaches the highest frequencies out of Africa. In fact, it represents about 85% of the European E-M78 chromosomes with a clinal pattern of frequency distribution from the southern Balkan peninsula (19.6%) to western Europe (2.5%). The same haplogroup is also present at lower frequencies in Anatolia (3.8%), the Near East (2.0%), and the Caucasus (1.8%). In Africa, haplogroup E-V13 is rare, being observed only in northern Africa at a low frequency (0.9%). The European E-V13 microsatellite haplotypes are related to each other to form a nearly perfect star-like network (fig. 4*A*), a likely consequence of a rapid demographic expansion (Jobling et al. 2004). The TMRCA of the European E-V13 chromosomes turns out to be 4.0–4.7 ky (under 2 different demographic expansion scenarios, see Subjects and Methods; 95% CI 3.5–4.6 ky and 4.1–5.3 ky, respectively). On the other hand, when only E-V13 chromosomes from western Asia are considered, the resulting network (fig. 4*B*) does not show such a star-like shape, and a much earlier TMRCA of 11.5 ky (95% CI 6.8–17.0; fig. 1) is obtained. These results open the possibility of recognizing time windows for 1) population movements from the E-M78 homeland in northeastern Africa to Eurasia and 2) population movements from western Asia into Europe and later within Europe.

The low E-V13 frequency (0.9%) and microsatellite variance (0.13) in northern Africa do not support an antiquity greater than in western Asia. Thus, the most parsimonious and plausible scenario is that E-V13 originated in western Asia about 11 ky ago, and its presence in northern Africa is the result of a more recent introgression. Under this hypothesis, E-V13 chromosomes sampled in western Asia and their coalescence estimate detect a likely Paleolithic exit out of Africa of E-M78 chromosomes devoid of the V13 mutation, which later occurred somewhere in the Near East/Anatolia. The refinement of location for the source area of such movements and associated chronologies here attained may be relevant to controversies on the spread of cultures (and languages) between Africa and Asia in the corresponding time frames (Bellwood 2004; Ehret et al. 2004, and references therein).

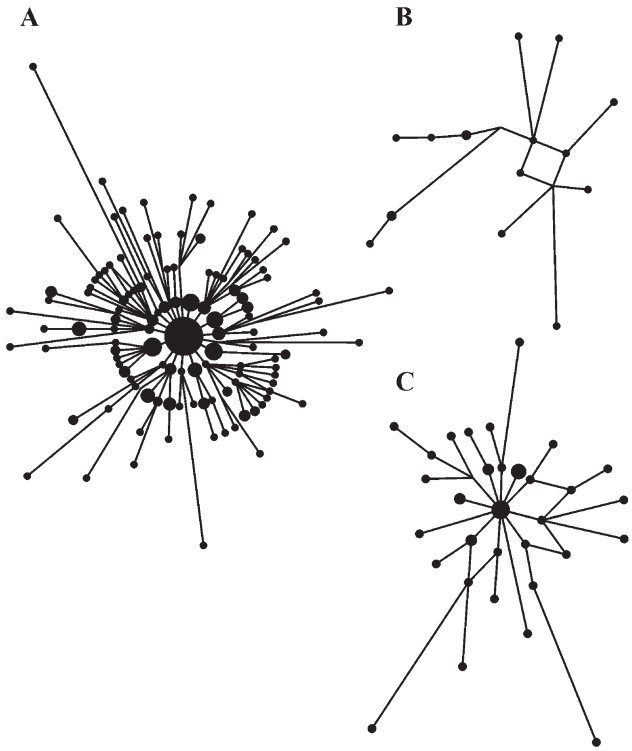


Fig. 4.—Microsatellite networks of haplogroups E-V13 in Europe (*A*), E-V13 in western Asia (*B*), and J-M12 in Europe (*C*).

As to a western Asia–Europe connection, our data suggest that western Asians carrying E-V13 may have reached the Balkans anytime after 17.0 ky ago, but expanded into Europe not earlier than 5.3 ky ago. Accordingly, the allele frequency peak is located in Europe, whereas the distribution of microsatellite allele variance shows a maximum in western Asia (fig. 5). Based on previously published data (Scozzari et al. 2001; Di Giacomo et al. 2004; Semino et al. 2004; Marjanovic et al. 2005), we observed that another haplogroup, J-M12, shows a frequency distribution within Europe similar to that observed for E-V13. In order to evaluate whether the present distribution of these 2 haplogroups can be the consequence of the same expansion/dispersal microevolutionary event, we first compared the 2 frequency distributions in Europe (J-M12 frequencies obtained from both published and new data; supplementary table 2, Supplementary Material online). We observed a high and statistically significant correspondence between the frequencies of the 2 haplogroups ($r = 0.84$, 95% CI 0.70–0.92). A similar result ($r = 0.85$, 95% CI 0.70–0.93) was obtained when the series was enlarged with the J-M12 data from Bosnia, Croatia, and Serbia (Marjanovic et al. 2005) matched with the frequencies of E-M78 cluster α (Peričić et al. 2005) as a proxy for haplogroup E-V13

(Cruciani et al. 2006). We then constructed a microsatellite network of 43 European J-M12 chromosomes (supplementary table 3, Supplementary Material online) and found a clear star-like structure (fig. 4*C*), a further feature shared with E-V13. This similarity was mirrored by a unimodal distribution of haplotype pairwise differences for both haplogroups (not shown). Finally, we used tetranucleotide microsatellite data in order to obtain a coalescence estimate for the J-M12 haplogroup in Europe. By taking into consideration 2 different demographic expansion models (see Subjects and Methods), we obtained TMRCA estimates very close to those of E-V13, that is, 4.1 ky (95% CI 2.8–5.4 ky) and 4.7 ky (95% CI 3.3–6.4 ky), respectively. Thus, the congruence between frequency distributions, shape of the networks, pairwise haplotypic differences, and coalescent estimates points to a single evolutionary event at the basis of the distribution of haplogroups E-V13 and J-M12 within Europe, a finding never appreciated before. These 2 haplogroups account for more than one-fourth of the chromosomes currently found in the southern Balkans, underlining the strong demographic impact of the expansion in the area.

Either environmental or cultural transitions are usually considered to be at the basis of dramatic changes of the size of human populations (Jobling et al. 2004). At least 4 major demographic events have been envisioned for this geographic area, that is, the post-Last Glacial Maximum expansion (about 20 kya) (Taberlet et al. 1998; Hewitt 2000), the Younger Dryas–Holocene reexpansion (about 12 kya), the population growth associated with the introduction of agricultural practices (about 8 kya) (Ammerman and Cavalli-Sforza, 1984), and the development of Bronze technology (about 5 kya) (Childe 1957; Piggott 1965; Renfrew 1979; Kristiansen 1998). Though large, the CI for the coalescence of both haplogroups E-V13 and J-M12 in Europe exclude the expansions following the Last Glacial Maximum or the Younger Dryas. Our estimated coalescence age of about 4.5 ky for haplogroups E-V13 and J-M12 in Europe (and their CIs) would also exclude a demographic expansion associated with the introduction of agriculture from Anatolia and would place this event at the beginning of the Balkan Bronze Age, a period that saw strong demographic changes as clearly testified from archeological records (Childe 1957; Piggott 1965; Kristiansen 1998). The arrangement of E-V13 (fig. 2*D*) and J-M12 (not shown) frequency surfaces appears to fit the expectations for a range expansion in an already populated territory (Klopfstein et al. 2006). Moreover, similarly to the results reported by Peričić et al. (2005) for E-M78 network α, the dispersion of E-V13 and J-M12 haplogroups seems to have mainly followed the river waterways connecting the southern Balkans to north-central Europe, a route that had already hastened by a factor 4-6 the spread of the Neolithic to the rest of the continent (Tringham 2000; Davison et al. 2006). This axis also served as a major route for the following millennia, enabling cultural and material (and possibly genetic) exchanges to and from central Europe (Childe 1957; Piggott 1965; Kristiansen 1998). Thus, the present work discloses a further level of complexity in the interpretation of the genetic landscape of southeastern Europe, this being to a large extent the consequence of a recent population increase in situ rather than the result of a mere flow of western Asian migrants in the early Neolithic. Indeed, Y-chromosomal data from regions to the north (Kasperavičiūtė et al. 2004), northwest (Luca et al. 2007), and west (Di Giacomo et al. 2004) to the Balkans show signatures of demographic events that match archeologically documented changes in the population size in the 1st millennia BC.

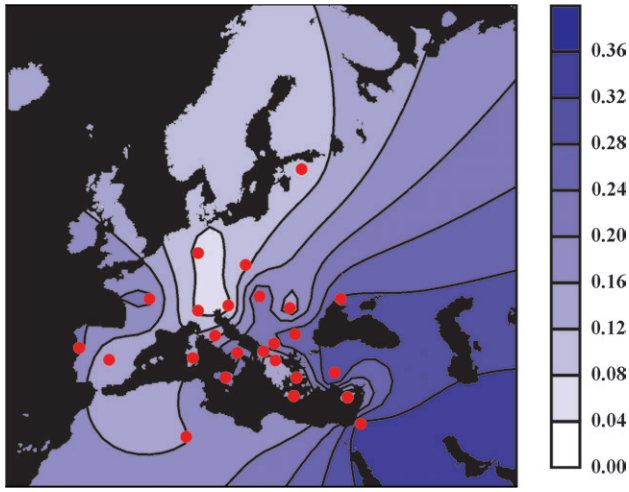


Fig. 5.—Mean variance spatial distribution of the Y-chromosome haplogroup E-V13 after pooling data from locations with <3 observations (see Subjects and Methods).

Concluding Remarks

The buildup of the present day male-specific Y-chromosome (MSY) diversity can be viewed as an increase of complexity due to the repeated addition of new variation to the preexisting background by 2 main mechanisms: immigration of differentiated MSY copies from outer regions and accumulation of novel MSY variants generated by new mutations in loco. The question is whether a DNA polymorphism, which is able to mark a specific episode indeed exists and is known. Recently, Sengupta et al. (2006) pointed out that combining high resolved phylogenetic hierarchy, haplogroup internal diversification, geography, and expansion time estimates can lead to the appropriate diachronic partition of the MSY pool. The DNA content of the MSY ensures that abundant diversity exists to proceed a long way in this process of phylogeographic refinement eventually leading to a level of resolution for human history comparable with, or even greater, than that achieved by mitochondrial DNA (Torroni et al. 2006).

Supplementary Material

Supplementary figure 1 and tables 1–3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to all the donors for providing DNA samples and to the people that contributed to the sample

collection. In particular, we thank João Lavinha (for the Portuguese samples); Farha El Chennawi, Anne Cambon-Thomsen, M.S. Issad, Eric Crubézy, Abdellatif Baali, Mohammed Cherkaoui, and Mohammed Melhaoui for their help in the collection of the Moroccan, Algerian, and Egyptian Berbers samples; and the National Laboratory for the Genetics of Israeli Populations. The useful comments and suggestions of 2 anonymous reviewers are gratefully acknowledged. This research received support from Grandi Progetti Ateneo, Sapienza Università di Roma (to R.S.), and the Italian Ministry of the University (Progetti di Ricerca di Interesse Nazionale 2005 to R.S. and Fondo Integrativo Speciale Ricerca 1999 to G.B. and R.S.). Russian samples were collected in the frame of an Italian-Russian scientific-technological project (3.RB3). The sampling of the Berbers was made within the framework of the Inserm ⟨⟨Réseau Nord/Sud⟩⟩ N°490NS1 (Mozabite Berbers), "The Origin of Man, Language and Languages", EUROCORES Programme and benefited from funding by the Région Midi-Pyrénées (Toulouse, France), the CNRS, and the E.C. Sixth Framework Programme under Contract ERASCT-2003-980409.

Literature Cited

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

Ammerman AJ, Cavalli-Sforza LL. 1984. The neolithic transition and the genetics of populations in Europe. Princeton: Princeton University Press.

Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, Makrelouf M, Pascali VL, Novelletto A, Tyler-Smith C. 2004. A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. Am J Hum Genet. 75:338–345.

Bandelt H-J, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 16:37–48.

Bandelt H-J, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. Genetics. 141:743–753.

Behar DM, Garrigan D, Kaplan ME, Mobasher Z, Rosengarten D, Karafet TM, Quintana-Murci L, Ostrer H, Skorecki K, Hammer MF. 2004. Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. Hum Genet. 114:354–365.

Bellwood P. 2004. The origins of Afroasiatic. Science 306:1681.

Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, 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.

Bosch E, Calafell F, González-Neira A, et al. (13 co-authors). 2006. Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. Ann Hum Genet. 70:459–487.

Butler JM, Schoske R, Vallone PM, Kline MC, Redd AJ, Hammer MF. 2002. A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. Forensic Sci Int. 129:10–24.

Cann HM, de Toma C, Cazes L, et al. (41 co-authors). 2002. A human genome diversity cell line panel. Science. 296:261–262.

Childe VG. 1957. The dawn of European civilization. London: Routledge and Kegan Paul.

Cinnioğlu C, King R, Kivisild T, et al. (15 co-authors). 2004. Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet. 114:127–148.

Cressie NAC. 1991. Statistics for spatial data. New York: John Wiley and Sons Inc.

Cruciani F, La Fratta R, Santolamazza P, et al. (19 co-authors). 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.

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. 27:831–832.

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.

Davison K, Dolukhanov P, Sarson GR, Shukurov A. 2006. The role of waterways in the spread of the Neolithic. J Archaeol Sci. 33:641–652.

Di Giacomo F, Luca F, Popa LO, et al. (27 co-authors). 2004. Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet. 115:357–371.

Ehret C, Keita SOY, Newman P. 2004. The origins of Afroasiatic. Science. 306:1680–1681.

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.

Flores C, Maca-Meyer N, González AM, Oefner PJ, Shen P, Pérez JA, Rojas A, Larruga JM, Underhill PA. 2004. Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: implications for population demography. Eur J Hum Genet. 12:855–863.

Flores C, Maca-Meyer N, Larruga JM, Cabrera VM, Karadsheh N, Gonzalez AM. 2005. Isolates in a corridor of migrations: a high-resolution analysis of Y-chromosome variation in Jordan. J Hum Genet. 50:435–441.

Forster P, Harding R, Torroni A, Bandelt H-J. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet. 59:935–945.

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

Gonçalves R, Freitas A, Branco M, Rosa A, Fernandes AT, Zhivotovsky LA, Underhill PA, Kivisild T, Brehm A. 2005. Y-chromosome lineages from Portugal, Madeira and Açores record elements of Sephardim and Berber ancestry. Ann Hum Genet. 69:443–454.

Gusmão L, Sánchez-Diz P, Calafell F, et al. (42 co-authors). 2005. Mutation rates at Y chromosome specific microsatellites. Hum Mutat. 26:520–528.

Hammer MF, Blackmer F, Garrigan D, Nachman MW, Wilder JA. 2003. Human population structure and its effects on sampling Y chromosome sequence variation. Genetics. 164:1495–1509.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. Nature. 405:907–913.

Jobling MA, Hurles ME, Tyler-Smith C. 2004. Human evolutionary genetics. New York: Garland Science.

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

Kasperavičiūtė D, Kučinskas V, Stoneking M. 2004. Y chromosome and mitochondrial DNA variation in Lithuanians. Ann Hum Genet. 68:438–452.

Kayser M, Brauer S, Cordaux R, et al. (15 co-authors). 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Mol Biol Evol. 23:2234–2244.

Klopfstein S, Currat M, Excoffier L. 2006. The fate of mutations surfing on the wave of a range expansion. Mol Biol Evol. 23:482–490.

Krings M, Salem AEH, Bauer K, et al. (13 co-authors). 1999. mtDNA analysis of Nile River Valley populations: a genetic corridor or a barrier to migration? Am J Hum Genet. 64:1166–1176.

Kristiansen K. 1998. Europe before history. Cambridge: Cambridge University Press.

Luca F, Di Giacomo F, Benincasa T, Popa LO, Banyko J, Kracmarova A, Malaspina P, Novelletto A, Brdicka R. 2007. Y-chromosomal variation in the Czech Republic. Am J Phys Anthropol. 132:132–139.

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. 74:532–544.

Malaspina P, Cruciani F, Ciminelli BM, et al. (24 co-authors). 1998. Network analyses of Y-chromosomal types in Europe, northern Africa, and western Asia reveal specific patterns of geographic distribution. Am J Hum Genet. 63:847–860.

Malaspina P, Cruciani F, Santolamazza P, et al. (24 co-authors). 2000. Patterns of male-specific inter-population divergence in Europe, West Asia and North Africa. Ann Hum Genet 64: 395–412.

Marjanovic D, Fornarino S, Montagna S, et al. (14 co-authors). 2005. The peopling of modern Bosnia-Herzegovina: Y-chromosome haplogroups in the three main ethnic groups. Ann Hum Genet. 69:757–763.

Mohyuddin A, Ayub Q, Underhill PA, Tyler-Smith C, Mehdi SQ. 2006. Detection of novel Y SNPs provides further insights into Y chromosomal variation in Pakistan. J Hum Genet. 51:375–378.

Olivieri A, Achilli A, Pala M, et al. (15 co-authors). 2006. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science. 314:1767–1770.

Peričić M, Barać Lauc L, Martinović Klarić I, et al. (18 co-authors). 2005. High-resolution phylogenetic analysis of southeastern Europe traces major episodes of paternal gene flow among Slavic populations. Mol Biol Evol. 22:1964–1975.

Piggott S. 1965. Ancient Europe from the beginnings of agriculture to classical antiquity. Edinburgh, UK: Edinburgh University Press.

Regueiro M, Cadenas AM, Gayden T, Underhill PA, Herrera RJ. 2006. Iran: tricontinental nexus for Y-chromosome driven migration. Hum Hered. 61:132–143.

Renfrew C. 1979. Before civilization. The radiocarbon revolution and prehistoric Europe. Cambridge: Cambridge University Press.

Rootsi S, Magri C, Kivisild T, et al. (45 co-authors). 2004. Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. Am J Hum Genet. 75:128–137.

Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet. 67:718–726.

Sanchez JJ, Hallenberg C, Børsting C, Hernandez A, Morling N. 2005. High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males. Eur J Hum Genet. 13:856–866.

Scozzari R, Cruciani F, Pangrazio A, et al. (17 co-authors). 2001. Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. Hum Immunol. 62:871–884.

Scozzari R, Cruciani F, Santolamazza P, et al. (17 co-authors). 1999. Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. Am J Hum Genet. 65:829–846.

Semino O, Magri C, Benuzzi G, et al. (16 co-authors). 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.

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.

Sengupta S, Zhivotovsky LA, King R, et al. (15 co-authors). 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.

Shen P, Lavi T, Kivisild T, et al. (11 co-authors). 2004. Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. Hum Mutat. 24:248–260.

Shen P, Wang F, Underhill PA, et al. (13 co-authors). 2000. Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci USA. 97:7354–7359.

Sims LM, Garvey D, Ballantyne J. 2007. Sub-populations within the major European and African derived haplogroups R1b3 and E3a are differentiated by previously phylogenetically undefined Y-SNPs. Hum Mutat. 28:97.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson J-F. 1998. Comparative phylogeography and postglacial colonization routes in Europe. Mol Ecol. 7:453–464.

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

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

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

Tringham R. 2000. Southeastern Europe in the transition to agriculture in Europe: bridge, buffer, or mosaic. In: Price TD, editor. Europe's first farmers. Cambridge: Cambridge University Press. p. 19–56.

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, et al. (21 co-authors). 2000. Y chromosome sequence variation and the history of human populations. Nat Genet. 26:358–361.

Wells RS, Yuldasheva N, Ruzibakiev R, et al. (27 co-authors). 2001. The Eurasian heartland: a continental perspective on Y-chromosome diversity. Proc Natl Acad Sci USA. 98:10244–10249.

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. 36:1122–1125.

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. (17 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.

Zhivotovsky LA, Underhill PA, Feldman MW. 2006. Difference between evolutionarily effective and germ-line mutation rate due to stochastically varying haplogroup size. Mol Biol Evol. 23:2268–2270.

Sarah Tishkoff, Associate Editor

Accepted March 4, 2007

CHAPTER 5

AIM OF THE WORK

This project is focused on the study of interspecific and interindividual (in human) variation of the AKR7A2 gene (aldo-keto reductase family 7, member A2).

THE MAIN AIM IS TO EXPLORE THE DEGREE OF VARIATION IN THE GENE AND TO TEST THE HYPOTHESIS THAT THE CURRENT PATTERN OF VARIATION WAS SHAPED BY NATURAL SELECTION.

The rationale for the choice of the AKR7A2 gene is the following:

1. AKR7A2 has been shown to be responsible for the synthesis of GHB in neural cells. Given the relevant biological effects of GHB at the cellular, tissutal and organismal livels, AKR7A2 is a candidate as target of selection on a variety of phenotypes;

2. other works in this laboratory explored the interspecific (Blasi et al. 2006) and intraspecific (Leone et al. 2006) variation of another gene encoding an enzyme of the same metabolic pathway, i.e. SSADH or NAD+-dependent succinic semialdehyde dehydrogenase. The two enzymes (SSADH and AKR7A2) in fact metabolize the same substrate, i.e. succinic semialdehyde produced in the catabolism of the neurotransmitter gamma-amino butyrate (GABA).

In other to reach the main aim, both original experimental data and data resulting from genome-scale projects and available on the web were used, in four phases of the research that addressed four intermediate aims.

First, a brief recognition of the evolution of the gene AKR7A2 in the human lineage in comparison to other mammals is performed, taking advantage of the recently released genomic sequences for a number of species and a number of publicly available computer programs for analysis. This because an increasing number of works in the literature show that evolutionary

trends and selection pattern observed at the inter-specific level are often replicated at the intra-specific level. The intermediate aim is to evaluate selection relaxation in the human lineage.

Second, a preliminary exploration of variation was performed by resequencing exon two in a limited sample of subjects. This intermediate aim is a cost-effective evaluation of a large genotyping effort.

Third, SNPs which turned-out to be informative were genotyped in many populations representative of all continents.

Fourth, in order to confirm our results, we repeated the same analysis using the HapMap Phase II data for 8 SNPs spanning approximately 10 kb centered on the AKR7A2 coding region (International HapMap Consortium 2007).

Analyses at the intraspecific level are performed with the aim of exploring whether polymorphisms in genes contributing to the same metabolic pathway show overlapping features in their population distributions and evolutionary signatures.

CHAPTER 6

IDENTIFYING THE TARGET GENE

The gene aldo-keto reductase 7A2 (AKR7A2) codes for an enzyme of the family of the aldo-keto reductases, isoenzymes that catalyse the NADPH-dependent oxidation and\or reduction of alcohol- and carbonyl-containing compounds (Kelly et al. 2002). AKR7A2 was first identified as a Aflatoxin B1 aldehyde reductase (AFB1) (Schaller et al. 1999).

Several aldo-keto reductases (AKRs), such as AKR7A2, are involved in the detoxification and metabolism of a variety of endogenous aldehydes and ketones. AKR7A2 catalyzes the NADPH-dependent reduction of two important classes of aldehydes and thus serves at least two different functions: whereas the reduction of aflatoxin B1 aldehyde is thought to protect against the toxicity and mutagenic potential of the potent aflatoxin

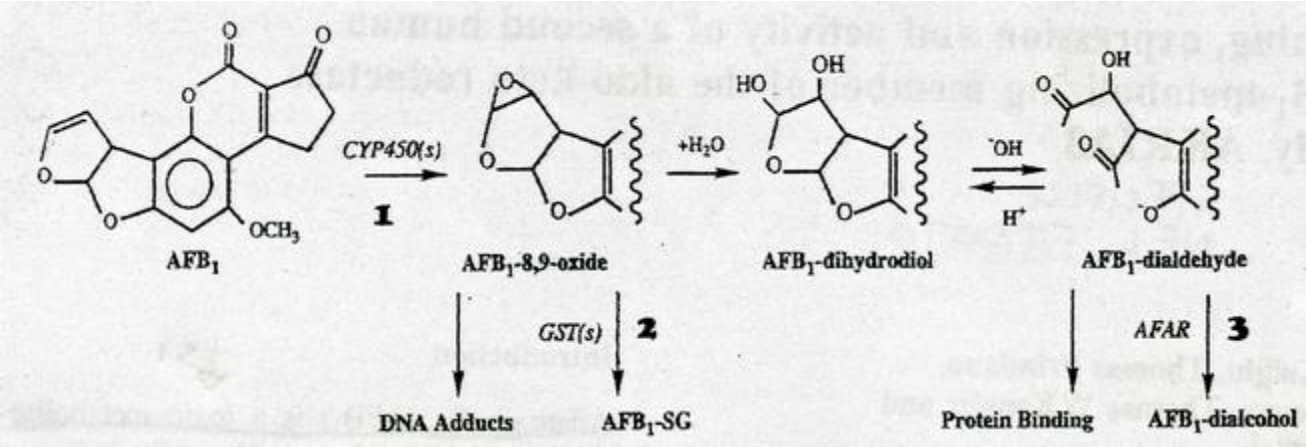
hepatocarcinogen, the AKR7A2 mediated reduction of the GABA metabolite succinic semialdehyde leads to the neuromodulatory compound gamma-hydroxybutyrate (GHB) in the central nervous system. The latter pathway, with accumulation of GHB, appears to be the reason for the observed neuropathological features (psychomotor retardation, language delay, seizures, hypotonia, ataxia) in succinic semialdehyde dehydrogenase deficiency (Pearl et al.2003).

Though we are interested here in the role of AKR7A2 in GHB production, its relevance in aflatoxin metabolism is discussed in Box1 and Box2.

Box1 Aflatoxin B1 is a mycotoxin. The term mycotoxin includes numerous secondary metabolites produced under conditions of high heat and humidity by microscopic and filamentous fungi, better known as molds; only a reduced subset of these manufacturing microrganisms can activate the secondary metabolic pathway that conduct to the synthesis of mycotoxin.
The mycotoxins are very different among them from the chemical point of view, and they show a notable range of biological effects, due to their ability to interact with different target organs and systems (Hsieh et al. 1987). For such reason they are classified in immunotoxins, hepatotoxins, neurotoxins or on the base of their chronic effect in mutagen and carcinogenic (Krogh et al. 1974).
All these biological activities are due to interactions of the mycotoxin and to their by-products with DNA, RNA, functional proteins, enzymatic and constituent cofactors of membrane.
The Aflatoxins are mycotoxin produced by fungi belonging to the class of the Ascomiceti, genus Aspergillus. In favourable environmental conditions of heat and damp, the spores of the Aspergillus germinate easily colonizing various substrates, among which cereals, peanuts, seeds oleaginous, corn and hay. Numerous types of Aflatoxin exist but that recognized more dangerous is the B1.
In certain developing regions of Africa and Asia, where food storage conditions are inadequate, humans are exposed to high levels of AFB1in their diet. Epidemiological evidences point out that in these populations this toxin contributes notably in the high incidence of the liver cancer. In European and north Americans countries such relationship has not been underlined as much the population is less exposed to this aflatoxin (Knight et al. 1999).

Box2 ***Liver metabolism of AFB1***

The harmful effects of AFB1 are attributed to reactive metabolites of this mycotoxin that produces adduct macromolecules.



Such metabolism begins with a reaction of epoxidation catalyzed by the CYP450 (cytochrome p450), that converts AFB1 in the AFB1 exo-8,9oxide which is reactive towards the DNA. In the phase 2 the G-glutathione-S-transferases can conjugate the reaction of epoxide with the reduction of the glutathione, preventing so the formation of DNA adducts (which can lead to mutations). AKR7A2 is an of the Aflatoxin B1 aldehyde reductase (AFAR) isoenzymes. These enzymes can reduce the dialdehyde protein-binding form of aflatoxin B1 (AFB1) to the nonbinding AFB1 dialcohol.
Two non-allelic isoenzymes are known in the human, AKR7A2 and AKR7A3. Both represent a key step in the phase-two of the detoxification of AFB1. (Knight et al. 1999).

Metabolism and effects of GHB

Gamma-Hydroxybutyrate (GHB) is an endogenous metabolite synthesized in the brain. There is strong evidence to suggest that GHB has an important role as a neurotransmitter or neuromodulator.

The human aldo-keto reductase AKR7A2 has been proposed previously to catalyze the NADPH-dependent reduction of succinic semialdehyde (SSA) to GHB in human brain (Lyon et al. 2007). AKR7A2 is mainly expressed in the central nervous system, particularly in the glial cells. Here, the protein AKR7A2 is able to catalyze the conversion of succinic semialdehyde SSA in the neuroactive compound gamma-hydroxybutyrate (GHB), acting

therefore as SSA reductase. In turn, SSA is the primary metabolite of GABA, the main inhibitory neurotransmitter in the central nervous system. The glial cells reabsorb the GABA after the neurotransmission and they degrade it according to the scheme reported in fig.1:

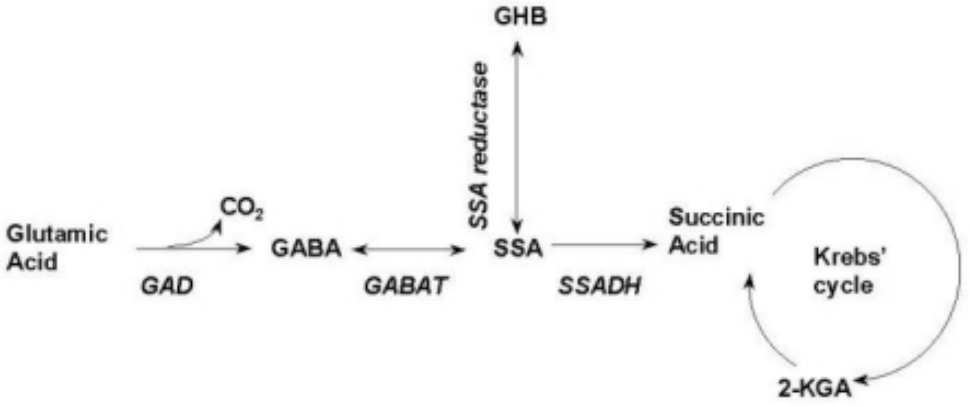


Fig.1 Schematic rapresentation of the GABA degradative pathway discussed in the text.

Two enzymes have been shown to catalyze the NADPH-dependent reduction of SSA in human brain. Both these enzymes have been identified as members of the aldo-keto reductase (AKR) family of enzymes: aldehyde reductase AKR1A1 and a dimeric SSA reductase AKR7A2, previously characterized as aflatoxin aldehyde reductase. AKR7A2 has a higher affinity for SSA than AKR1A1.

Not all of the SSA produced by transamination of GABA is reduced to GHB. SSA can also be oxidized by the mitochondrially located enzyme succinic semialdehyde dehydrogenase (SSADH) producing succinate for entry into the Krebs cycle. Defects in SSADH are found in individuals suffering from 4-hydroxybutyric aciduria, which results in an abnormal accumulation of GHB (Pearl et al. 2003). SSA that cannot be oxidized in these individuals is reduced to GHB by SSA reductase, and it is the high levels of GHB that are presumed to cause the clinical syndrome associated with SSADH deficiency, characterized by psychomotor retardation, hypotonia, ataxia, and poorly developed to absent speech.

It is also known, that the catabolite GHB has effects on the behaviour and perhaps also at cognitive level (Vasiliou et al 2004). It is believed that large increases in brain GHB concentration, following external administration, hyperstimulate the GHB receptors, and this mechanism is likely to be the basis of the major pharmacological effects of GHB.

Endogenous GHB is unevenly distributed in the central nervous system (CNS) as well as in peripheral organs. The significance of endogenous GHB, which shows most of the features of a neurotransmitter/neuromodulator, is suggested by clinical studies showing that dramatic increases of its endogenous concentrations, ensuing to mutations of enzymes involved in GABA metabolism, are associated with relevant alterations of cognitive and motor functions as well as with enhanced neuronal excitability.

GHB is also a drug, and when administered at pharmacological doses and depending on the dose, it produces sedation, euphoria, anxiolysis, hypnosis, and anesthesia. Thus, it has been proven efficacious in the induction of general anesthesia, in the pharmacotherapy of narcolepsy and in reducing the symptomatology of alcohol withdrawal and alcohol craving in alcoholics. GHB appears endowed with neuroprotective properties; as an example, in experimental models of transient global ischemia it reduces the brain tissue damage as well as the ensuing functional impairment. The use of GHB as a therapeutic tool has been thus far limited by the threat represented by its increasing abuse, both as a recreational drug and for its purported anabolic effects. In addition, the molecular mechanisms underlying its effects on such a variety of functional aspects are still controversial. Preclinical evidence shows that its numerous actions may depend on the activation of specific GHB receptors (GHB-r) and /or GABAB receptors, for whose, however, GHB shows a rather low affinity.

CHAPTER 7

MATERIALS AND METHODS

Genomic organization of AKR7A2 gene

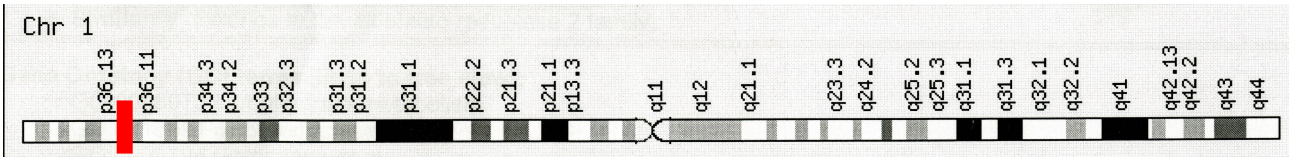
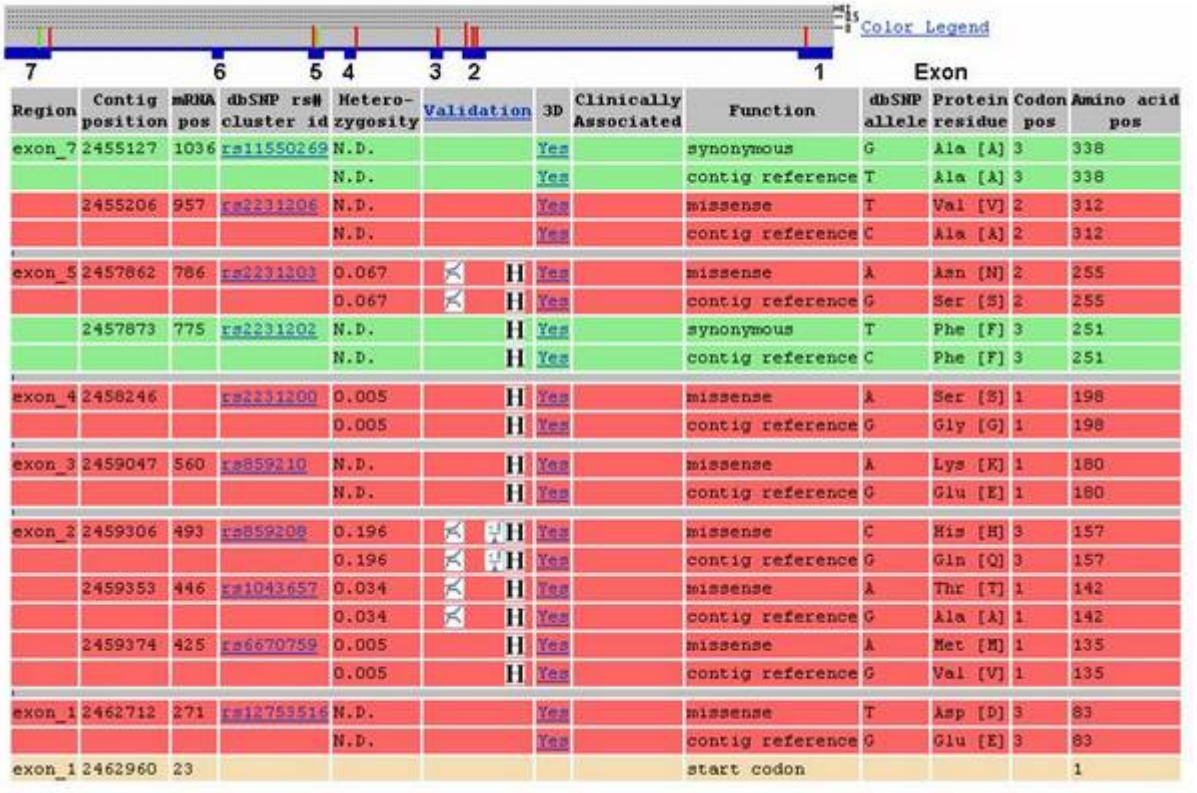


Fig.2 Chromosome1. Position of the AKR7A2 gene is reported in red.

The gene AKR7A2 is located on the short arm of chromosome 1 (Praml et al. 2003), at chromosome positions 19375768 to 19383945 in NCBI B35 assembly. It consists of seven exons. The references for the coding and genomic sequences are NM_003689.2 and AL035413, respectively.

A preliminary exploration of SNPs reported in dbSNP (www.ncbi.nml.gov/projects/SNP) was performed. This exploration revealed that exon 2 is enriched in nonsynonymous coding SNPs. This exon was then selected for subsequent experimental analyses.



Region	Contig position	mRNA pos	dbSNP rs# cluster id	Hetero-zygosity	Validation	3D	Clinically Associated	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
exon_7	2455127	1036	rs11550269	N.D.		Yes		synonymous	G	Ala [A]	3	338
				N.D.		Yes		contig reference	T	Ala [A]	3	338
	2455206	957	rs2231206	N.D.		Yes		missense	T	Val [V]	2	312
				N.D.		Yes		contig reference	C	Ala [A]	2	312
exon_5	2457862	786	rs2231203	0.067		Yes		missense	A	Asn [N]	2	255
				0.067		Yes		contig reference	G	Ser [S]	2	255
	2457873	775	rs2231202	N.D.		Yes		synonymous	T	Phe [F]	3	251
				N.D.		Yes		contig reference	C	Phe [F]	3	251
exon_4	2458246		rs2231200	0.005		Yes		missense	A	Ser [S]	1	198
				0.005		Yes		contig reference	G	Gly [G]	1	198
exon_3	2459047	560	rs859210	N.D.		Yes		missense	A	Lys [K]	1	180
				N.D.		Yes		contig reference	G	Glu [E]	1	180
exon_2	2459306	493	rs859208	0.196		Yes		missense	C	His [H]	3	157
				0.196		Yes		contig reference	G	Gln [Q]	3	157
	2459353	446	rs1043657	0.034		Yes		missense	A	Thr [T]	1	142
				0.034		Yes		contig reference	G	Ala [A]	1	142
	2459374	425	rs6670759	0.005		Yes		missense	A	Met [M]	1	135
				0.005		Yes		contig reference	G	Val [V]	1	135
exon_1	2462712	271	rs12753516	N.D.		Yes		missense	T	Asp [D]	3	83
				N.D.		Yes		contig reference	G	Glu [E]	3	83
exon_1	2462960	23						start codon				1

Fig. 3 list of coding SNPs in AKR7A2. Synonymous SNPs are reported in green, nonsynonymous SNPs are reported in red. A summary view of SNP locations in the gene is given at top. SNPs features are detailed.

Subjects. We typed the HDGP panel (Cann et al. 2002). This consists in a resource of 1064 cultured lymphoblastoid cell lines (LCLs) from 1051 individuals in 51 different world populations . These LCLs were collected from various laboratories by the HGDP and CEPH in order to provide unlimited supplies of DNA for studies of sequence diversity and history of modern human populations.

In addition, we also examined 147 subjects of Calabria from five different locations (Acri, Lamezia Terme, Lungro, Locri, Paola) and 6 Egyptian subjects collected in the frame of a screening program for hemoglobinopathies. DNA was extracted from venous blood in EDTA, blood drops adsorbed on paper or buccal smear. Informed consent was obtained verbally in all cases.

Experimental procedures. Two primers were designed, that allow the amplification of a 803 bp DNA segment, encompassing AKR7A2 exons 2 and 3 (positions 61100 to 61902 in AL035413).

AKR7A2g-1	5′-ttgaaggtttggaggagcctcac-3′
AKR7A2g-3	5′-gtgcctctgctctcatgag-3′

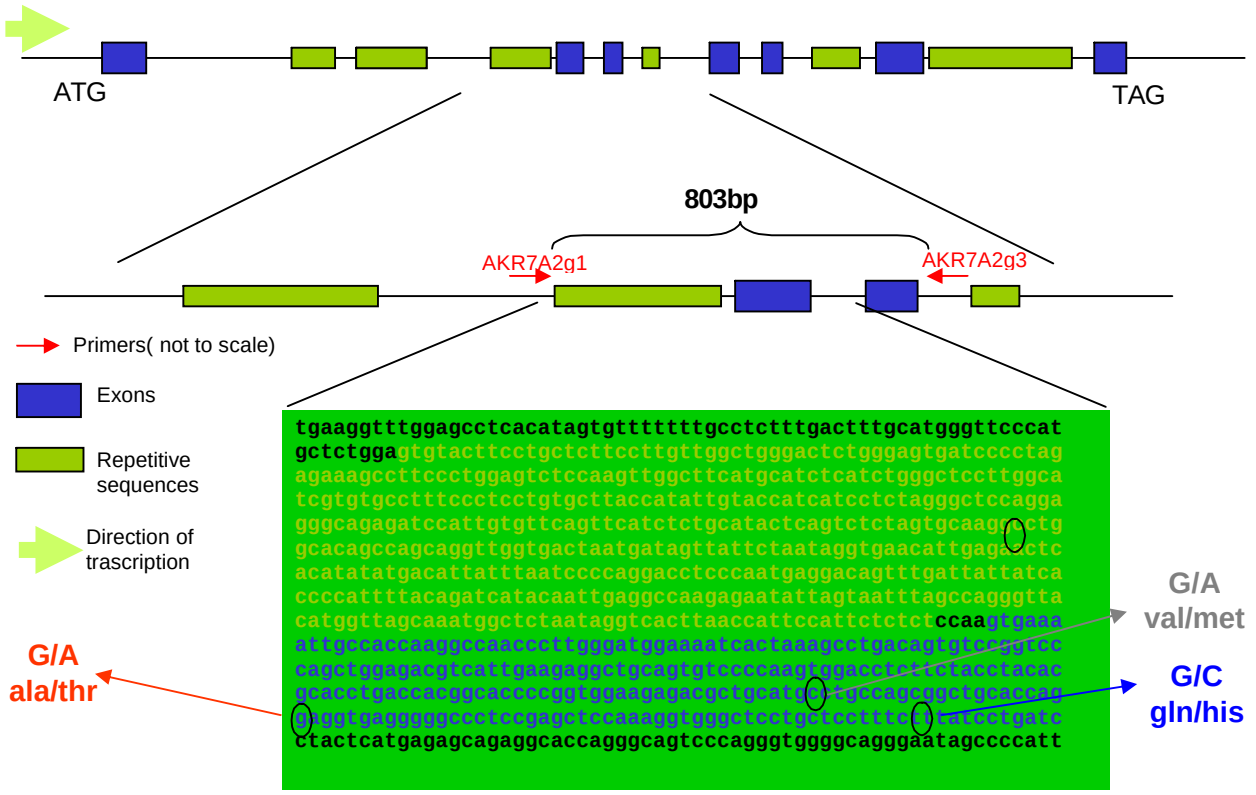


Fig. 4 Chromosome organization of the human AKR7A2 gene. Blu boxes represent exons. Green boxes represent repetitive sequences. A close-up of exons 2-3 is given in the second line. The sequence of the 803bp PCR product is reported at bottom. The locations of three coding nonsynonymous SNPs discussed in the text are shown, plus one intronic SNP (rs7525784).

The PCR product was used as template for the a) sequencing reaction, or b) for genotyping.

Total DNA was amplified in 25 µl reaction mixtures containing 200 ng of genomic DNA, 0.5U of Taq DNA polymerase, 0.2 mM dNTPs, 2 mM MgCl2 and 1X reaction buffer. Amplifications were performed in an Eppendorf thermal cycler. Cycling conditions were : predenaturation at 94°C for 5', denaturation at 94°C for 30'', annealing at 60°C for 30'', extension at 72°C for 1.30'', for 35 cycles, and a final extension at 72°C for 10'.

a) Sequencing

The amplified fragments were purified by QIAquick PCR purification Kit (Qiagen) and sequenced by using the BigDye Terminator Cycle Sequencing Ready Reaction Kit (Perkin Elmer) on an ABI 310 automated sequencer (Applied Biosystems).

Electropherograms were aligned with the reference sequence and visually inspected.

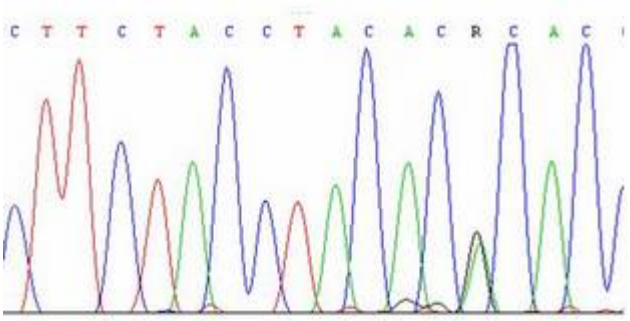


Fig. 5 portion of an electropherogram showing heterozygosity at rs1043657.

Sequences were aligned with the reference sequence AL035413. This part of the analysis was applied to 6 Calabrian and 6 Egyptian subjects.

b) Genotyping

Individual genotypes at positions rs859208 and rs1043657 and rs6670759 were determined with two independent methods.

In the first method, applied only to the Calabrian subjects, DNA was initially amplified as described above. 5 ul of the amplified product were digested with the enzyme BsrFI. A BsrFI site is generated in the presence of C at position 708 (rs859208) of the amplified product. The following restriction pattern is obtained:

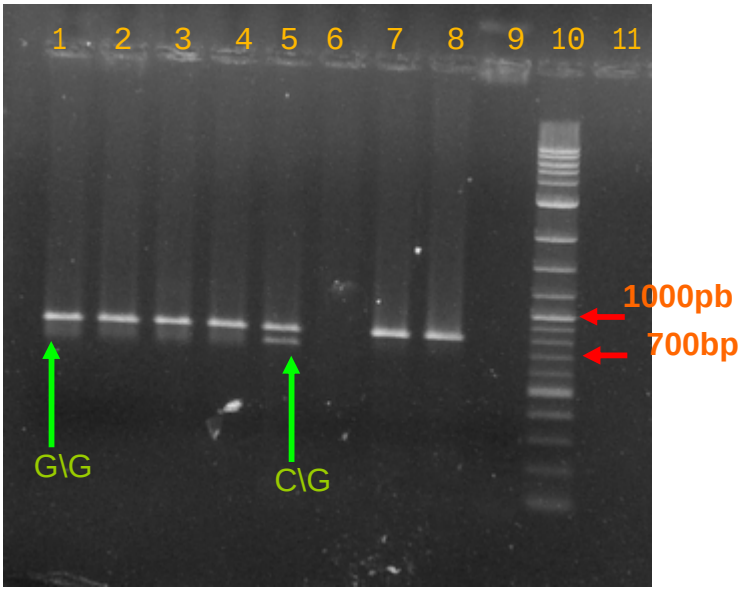


Fig 6. Restriction pattern and rs859208 genotypes:
Homozygote G/G 803bp
Heterozygote C/G 803bp, 708bp (lane 5)
(no homozygote C/C was found among the Calabrian subjects analyzed)
Negative controls are shown in lanes 6 and 9.
Molecular weight marker is shown in lane 10

In the second method, the PCR product was spotted on nylon membranes and hybridized with ^{32}P labeled allele-specific oligonucleotide (ASO) probes as described (Blasi et al. 2006). Table 1 reports hybridization conditions and allele status.

Table 1. ASO probes and conditions used in AKR7A2 genotyping (the variant position is underlined)

SNP	Primer name	Primer sequence 5′-3′	Allele state	Washing temp
rs859208	Rs859208-C	CCTGCCACCGGCTG	Ancestral	48°C
	Rs859208-G	CCTGCCAGCGGCTG	Derived	48°C
rs1043657	Rs1043657-G	ACCTACACGCACCTGAC	Ancestral	51°C
	Rs1043657-A	ACCTACACACACCTGAC	Derived	51°C
rs6070759	Rs6070759-G	TCCCCAAGTGGACCT	Ancestral	45°C
	Rs6070759-A	TCCCCAAATGGACCT	Derived	45°C

Each amplified product was spotted in duplicate and each duplicate was hybridized with probes that differ for a single nucleotide, corresponding to the genomic variation. As the three SNPs reside in the same PCR product, each membrane could be sequentially hybridized with ASO probes corresponding to different variant positions. The results obtained after two cycles of hybridization, and the corresponding genotypes are shown in figure 7.

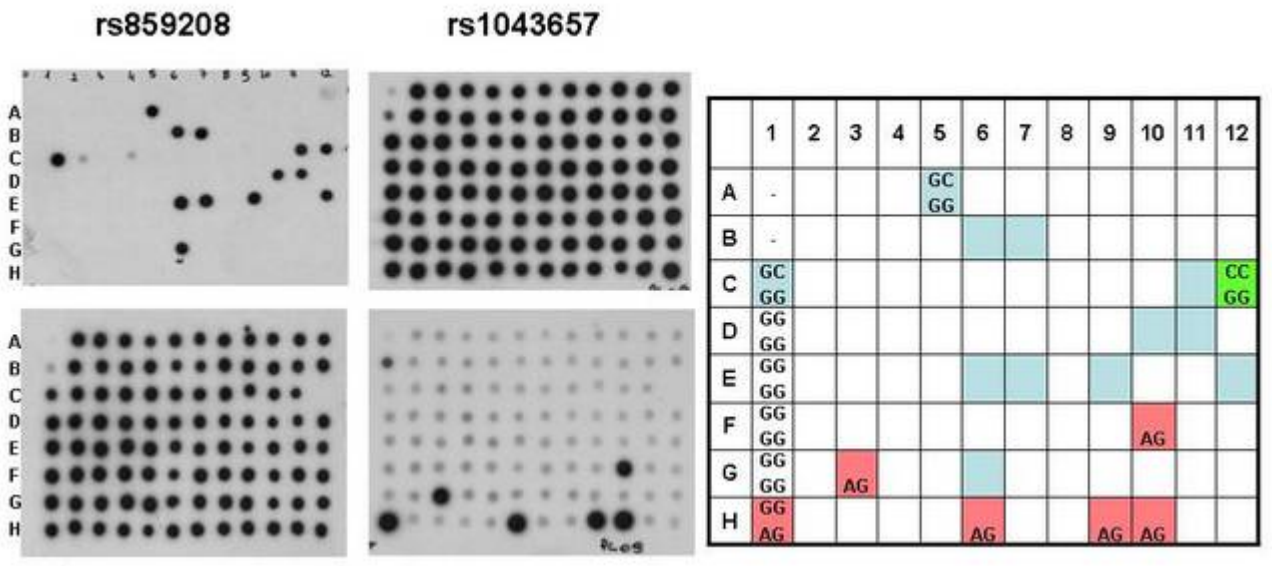


Fig. 7 Autoradiograph of nylon filters after hybridization with two ASO probes for the two SNPs indicated. Interpretation of genotypes is exemplified to the right. Position A1 and B1 are void as negative controls.

Data analysis.

AKR7A2 in the human lineage in comparison to other mammals

Complete cDNA sequences for the Macaque, Dog, Rat and Mouse orthologues of human AKR7A2 were downloaded from the homologene data base (www.ncbi.nlm.nih.gov) and aligned with the program ClustalX (Thompson et al. 1997). The Chimpanzee sequence was excluded, due to incomplete alignment at the time of analyses. Downloaded sequences were used to determine the ancestral and derived allele at each SNP. A likelihood ratio test for constancy of the non-synonymous vs. synonymous rates of nucleotide changes was performed as described by Yang and Nielsen (2000, 2002) with the program PAML (Yang. 1997). This program calculates the relationship among rate of non synonymous substitutions (Ka) and rate of synonymous substitutions (Ks) along the branches of a phylogenetic tree. It is used for testing the hypothesis of a constant rate among the branches. A likelihood ratio test is performed by comparing trees in which the ratio is allowed to vary across branches and that in which it is constant.

HapMap data were downloaded from www.hapmap.org as phased haplotypes in the European, Yoruban and Chinese+Japanese populations.

The Arlequin 2.000 package (Schneider et al. 2000) was used to calculate haplotype diversity, to test for departure from the Hardy-Weinberg equilibrium and to evaluate fixation indexes under the AMOVA scheme.

Gene-Haplotype diversity

This is equivalent to the expected heterozygosity for diploid date. It is defined as the probability to randomly choose two different haplotypes. Gene diversity and is estimated as

$$H = (n/n-1)*\left(1-\sum_{i=1}^{k} p_i^2\right)$$

where *n* is the number of gene copies in the sample, *K* is the number of haplotypes, and *pi* is the sample frequency of the *i*-th haplotype.

AMOVA

The Analysis of Molecular Variance approach used in Arlequin (AMOVA, Excoffier et al. 1992) is essentially similar to other approaches based on analyses of variance of gene frequencies, but it takes into account the number of mutations between molecular haplotypes (which first need to be evaluated).

By defining groups of populations, the user defines a particular genetic structure that will be tested. A hierarchical analysis of variance partitions the total variance into covariance components due to

- intra-individual differences,
- inter-individual differences,
- inter-population differences

The aim is to distinguish the components of the variability (measured as variance in the comparisons among all the possible couples of individuals) in three levels (fig.8):

• among individuals inside every population (σc^2)

• among different populations belonging to the same group of populations (cluster) (σb^2)

• among groups of populations (σa^2)

The total variance is the sum of the three components:

$$\sigma tot^2 = \sigma a^2 + \sigma b^2 + \sigma c^2$$

while the indexes of fixation are defined as:

$Fst = (\sigma a^2 + \sigma b^2) / (\sigma tot)$ measure the difference among the populations in comparison to the total one

$Fct = (\sigma a^2) / \sigma tot^2$ measures the difference among the groups of populations in comparison to the total one

Fsc = (σb^2)/ ($\sigma b^2 + \sigma c^2$) measure the difference among populations inside the groups of affiliation.

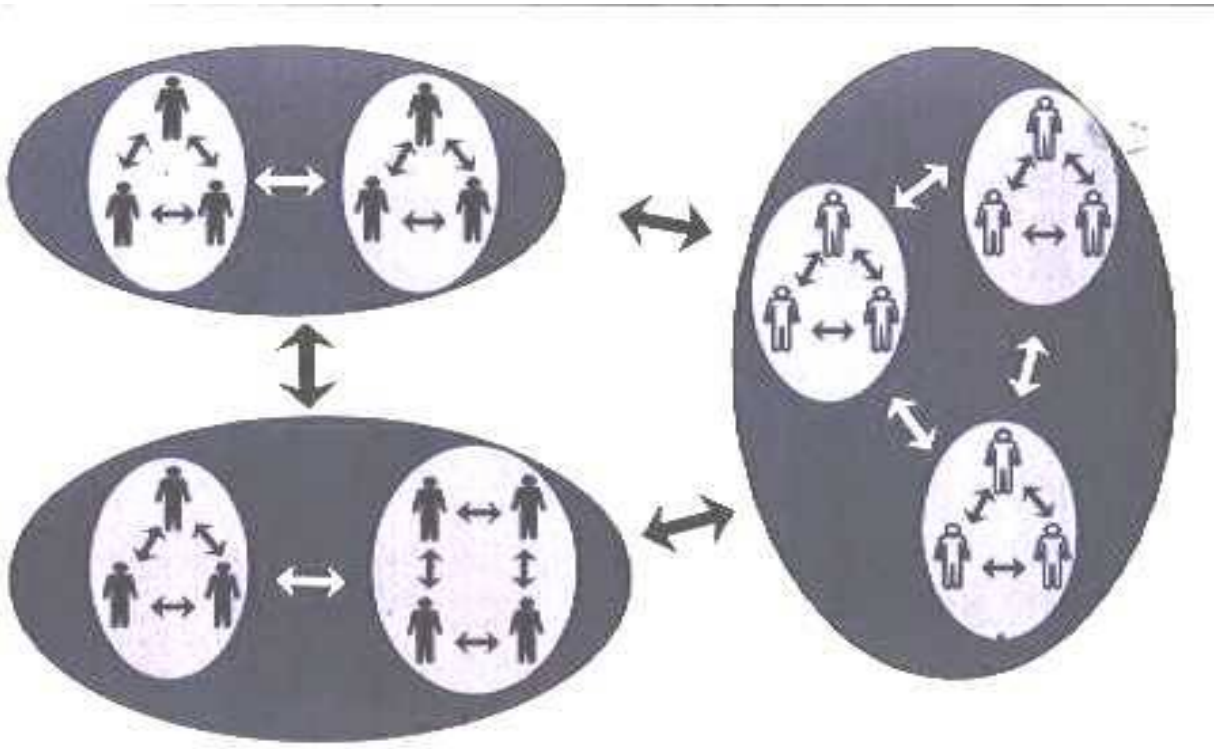


Fig.8 Representation of the hierarchical grouping of individuals applied in the analysis of the molecular variance. The black arrows inside the white ovals point out the comparisons among individuals within populations; the white arrows point out the comparisons among populations inside the clusters; the great black arrows point out the comparisons among clusters of populations.

In our analysis the population clusters corresponded to continental affilation i.e. Europe, Asia including Middle East, Africa, America and Oceania represented by PNG and Melanesia.

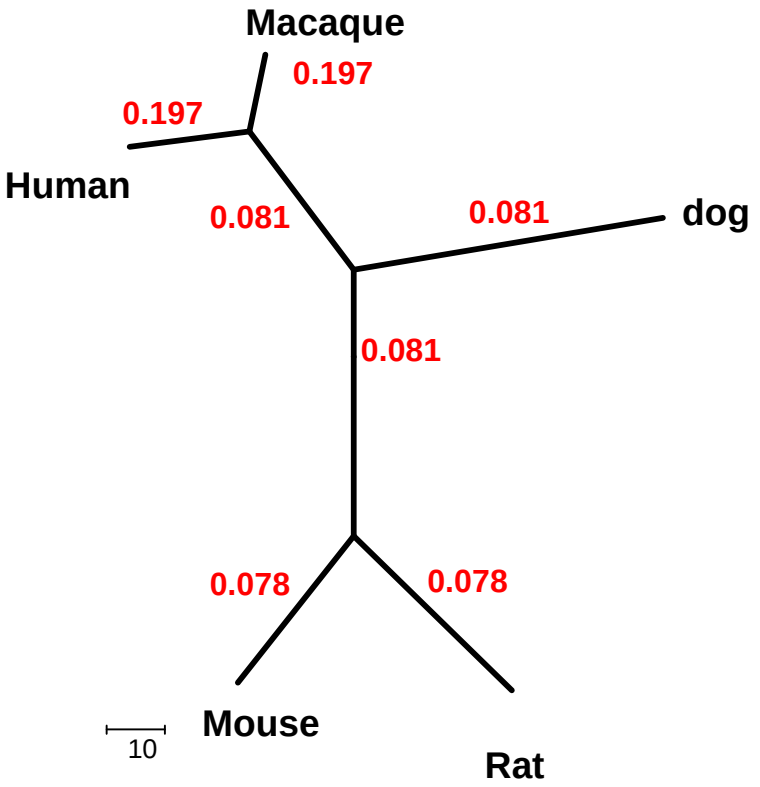
CHAPTER 9
RESULTS AND DISCUSSION

AKR7A2 in the human lineage in comparison to other mammals

We have analyzed the AKR7A2 phylogenetic tree as reconstructed in five mammalian species (human, macaque, mouse, rat and dog). The multiple alignment revealed that the chimpanzee sequence probably contains numerous errors in the form of insertions and deletions that in many cases alter the reading frame. It has then been excluded from the analysis.

As expected, the two primates clustered together, as well as the two rodents (fig. 9). We tested the hypothesis of three Ka/Ks ratios, for primates, rodents and the mammalian background. We obtained values of 0.197, 0.078 and 0.081, respectively. These results show an acceleration of the Ka/Ks ratio in primate lineage. These variations explained the data significantly better (χ^2=5.34, 1 g.d.l.) than a single Ka/Ks ratio.

When these values are compared with those obtained by Dorus et al. (2004) on genes expressed in the nervous system, the value for the rodents resulted within the normal range, whereas the value for the primates largely exceeded the average obtained by these authors (0.12 +/- 0.01). In conclusion the gene AKR7A2 shows an acceleration of the rate of aminoacid substitution even more pronounced than other genes for which selection for increased cerebral capacity and cognitive ability has been postulated.



$X^2=5.34$ $P<0.02$

Fig. 9 phylogenetic tree of the AKR7A2 gene in five mammalian species. Branch length is proportional to the number of substitutions. For each branch the Ka/Ks ratio is reported in red.

The inter-specific analysis reveals the ancestral state

The multiple alignement showed that for the SNP rs859208 the chimpanzee, the macaque, the dog and the mouse share a C, while the rat carries a A (fig.10).

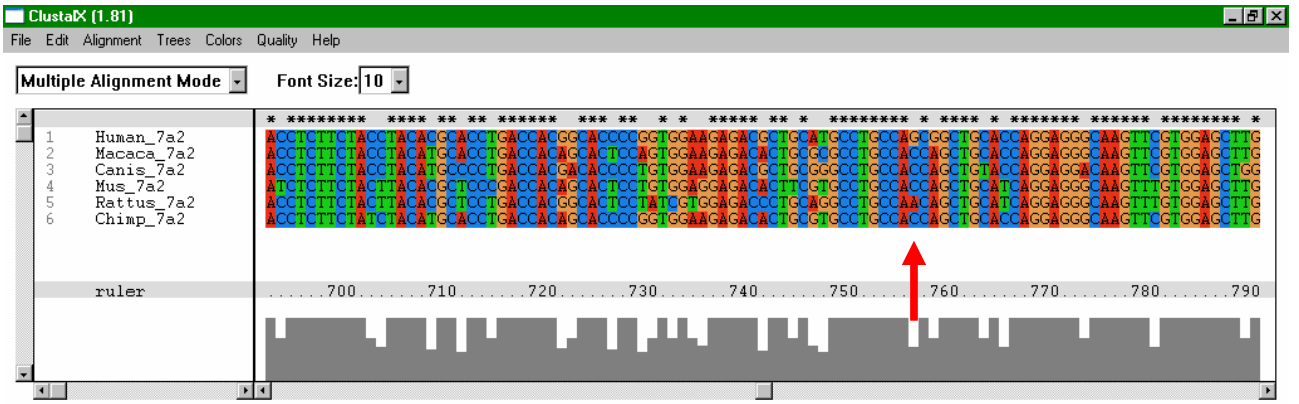


Fig.10 multiple alignement of AKR7A2 cDNA sequences in the region overlapping rs859208.

This result unambiguously identified the C as ancestral and the G as the derived allele. For the two other SNP, rs6670759 rs1043657 the ancestral alleles were identified as G and G, respectively, in agreement with dbSNP.

A preliminary exploration of human variation

In order to verify the polymorphism of AKR7A2 exon2, we sequenced the PCR product obtained in 12 subjects. One of the non-synonymous SNPs reported in the dbSNP (rs6670759, see fig. 4) did not show polymorphism in this small sample, while rs1043657, rs859208 and rs7525784 were polymorphic. Finally variation at a novel position was found in one subject. This was a G>C transversion in position 8 of exon 2.

The above results prompted us to test rs1043657 and rs859208 in the HGDP with a rapid assay (see materials and methods). We extended the typing to rs6670759 as this position resides within the same PCR product and requires only rehybridization of the same nylon membranes. The results are summarized in table 2.

rs6670759 turned out to be monomorphic in all populations. rs1043657 showed the highest frequency of the derived A allele especially in Europe and in north Africa (Algeria). The derived allele was not observed in south Africa, Far East, south-east Asia and Oceania.

rs859208 showed a completely different pattern: the ancestral C allele persists in Africa at high frequencies, reaching 0.75 in Pygmies. Within Africa the second highest frequencies is observed in Bantu-speaking populations (from Nigeria, Kenia and South-Africa). Lower frequencies are observed in Western and North-Africa. Outside Africa the derived G allele predominates. In Europe the ancestral allele C is rare. We typed 147 Calabrian subjects and found only one heterozygote. Another instance is observed among French. In western and central Asia the ancestral C is always below 0.05 , increasing to 0.09 in south east Asia and to >0.11 in Oceania. The derived G allele appears to be fixed in native Americans.

Table 2: results of analysis of frequencies in 52 population samples typed for rs859208, rs1043657 and rs6670759.

				rs859208					rs1043657					rs6670759
Continent	Geographic origin	Population	Sample origin	Sample size	Allele Der. G	Allele Anc. C	s.e.	H.W.	Sample size	Allele Der. A	Allele Anc. G	s.e.	H.W.	Sample size	Allele
Africa	Algeria	Mozabite	HGDP panel	30	0.883	0.116	0.058	1	30	0.100	0.900	0.055	1	30	1.000
	Senegal	Mandenka	HGDP panel	22	0.863	0.136	0.073	1	22	0.068	0.931	0.054	1	22	1.000
	Nigeria	Yoruba	HGDP panel	23	0.608	0.391	0.102	0.673	23		1.000		n.a.	23	1.000
	Central African R.	Biaka Pygmies	HGDP panel	31	0.274	0.725	0.080	1	33	0.090	0.909	0.050	1	33	1.000
	Dem. R. Congo	Mbuti Pygmies	HGDP panel	14	0.250	0.750	0.116	0.512	14		1.000		n.a.	14	1.000
	Kenia	Bantu N.E.	HGDP panel	12	0.666	0.333	0.136	1	12		1.000		n.a.	12	1.000
	Namibia	San	HGDP panel	7	0.928	0.071	0.097	1	7		1.000		n.a.	7	1.000
	South Africa	Bantu	HGDP panel	8	0.562	0.437	0.175	1	8		1.000		n.a.	8	1.000
Europe	France	French Basque	HGDP panel	5	1.000			n.a.	5	0.300	0.700	0.205	1	5	1.000
	France	French	HGDP panel	29	0.982	0.017	0.024	1	29	0.068	0.931	0.047	1	29	1.000
	Italy	Sardinian	HGDP panel	15	1.000			n.a.	15	0.066	0.933	0.064	1	15	1.000
	Italy	Bergamo	HGDP panel	9	1.000			n.a.	9	0.111	0.888	0.105	1	9	1.000
	Italy	Tuscan	HGDP panel	8	1.000			n.a.	8	0.125	0.875	0.117	1	8	1.000
	Italy	Southern Italia	THIS PAPER	147	0.993	0.007	0.007	n.a.	n.t.					n.t.	
	Orkney Islands	Orcadian	HGDP panel	16	1.000			n.a.	16	0.218	0.782	0.103	0.541	16	1.000
	Russia	Russian	HGDP panel	22	1.000			n.a.	22	0.250	0.750	0.092	1	22	1.000
	Russia	Adygei	HGDP panel	17	1.000			n.a.	17	0.088	0.911	0.069	1	17	1.000
Asia	Israel	Bedouin	HGDP panel	46	0.966	0.033	0.026	1	47	0.138	0.861	0.050	0.189	47	1.000
	Israel	Drusi	HGDP panel	45	0.977	0.022	0.022	1	45	0.122	0.877	0.049	0.108	45	1.000
	Israel	Palestinian	HGDP panel	24	0.979	0.020	0.029	1	24	0.104	0.895	0.062	1	24	1.000
	Pakistan	Brahui	HGDP panel	7	1.000			n.a.	7		1.000		n.a.	7	1.000
	Pakistan	Balochi	HGDP panel	23	0.978	0.021	0.030	1	23	0.021	0.978	0.030	1	23	1.000
	Pakistan	Hazara	HGDP panel	25	1.000			n.a.	25	0.040	0.960	0.039	1	25	1.000
	Pakistan	Makrani	HGDP panel	24	1.000			n.a.	24	0.041	0.958	0.040	1	24	1.000
	Pakistan	Sindhi	HGDP panel	25	0.980	0.020	0.028	1	25	0.060	0.940	0.047	1	25	1.000
	Pakistan	Pathan	HGDP panel	25	1.000			n.a.	25	0.080	0.920	0.054	0.12	25	1.000
	Pakistan	Kalash	HGDP panel	6	1.000			n.a.	6		1.000		n.a.	6	1.000
	Pakistan	Burusho	HGDP panel	2	1.000			n.a.	2		1.000		n.a.	2	1.000

					rs859208					rs1043657					rs6670759	
Continent	Geographic origin	Population	Sample origin	Sample size	Allele		s.e.	H.W.	Sample size	Allele		s.e.	H.W.	Sample size	Allele
					Der.	Anc.				Der.	Anc.				
					G	C				A	G				
	China	Tuja	HGDP panel	3	1.000			n.a.	3		1.000			3	1.000
	China	Yizu	HGDP panel	7	0.928	0.071	0.097	1	7		1.000			7	1.000
	China	Miazou	HGDP panel	10	0.850	0.150	0.113	1	10		1.000			10	1.000
	China	Orogen	HGDP panel	10	1.000			n.a.	10		1.000			10	1.000
	China	Daur	HGDP panel	10	1.000			n.a.	10		1.000			10	1.000
	China	Mongol	HGDP panel	10	0.950	0.050	0.069	1	10		1.000			10	1.000
	China	Hezhen	HGDP panel	10	1.000			n.a.	10		1.000			10	1.000
	China	Xibo	HGDP panel	9	1.000			n.a.	9		1.000			9	1.000
	China	Uighur	HGDP panel	9	1.000			n.a.	9	0.111	0.888	0.105		9	1.000
	China	Dai	HGDP panel	10	0.950	0.050	0.069	1	10		1.000			10	1.000
	China	Han	HGDP panel	38	1.000			n.a.	38		1.000			38	1.000
	China	Lau	HGDP panel	10	1.000			n.a.	10		1.000			10	1.000
	China	She	HGDP panel	10	1.000			n.a.	10		1.000			10	1.000
	China	naxi	HGDP panel	10	1.000			n.a.	10		1.000			10	1.000
	China	Tu	HGDP panel	10	1.000			n.a.	10		1.000			10	1.000
	Japan	Japanese	HGDP panel	31	0.967	0.032	0.032	1	31		1.000			31	1.000
	Cambodian	Cambodian	HGDP panel	11	0.909	0.090	0.086	1	11		1.000			11	1.000
	Siberia	Yakut	HGDP panel	25	0.980	0.020	0.028	1	25		1.000			25	1.000
Oceania	New Guinea	Papuan	HGDP panel	17	0.882	0.117	0.078	0.178	17		1.000			17	1.000
	Bougainville	NAN Melanesian	HGDP panel	22	0.863	0.135	0.073	1	22		1.000			22	1.000
America	Mexico	Pima	HGDP panel	25	1.000			n.a.	25	0.040	0.960	0.039		25	1.000
	Mexico	Maya	HGDP panel	25	1.000			n.a.	25		1.000			25	1.000
	Colombia	Colombian	HGDP panel	13	1.000			n.a.	13		1.000			13	1.000
	Brazil	Karitiana	HGDP panel	24	1.000			n.a.	24		1.000			24	1.000
	Brazil	Surui	HGDP panel	18	1.000			n.a.	18	0.027	0.972	0.038		18	1.000
Europe			Hap-Map Project	60	0.912	0.008	0.008		60		1.000				
Africa	Nigeria	Yoruba	Hap-Map Project	60	0.727	0.283	0.003		60	0.025	0.975	0.014			
Asia		Chinese+Japanes	Hap-Map Project	90	0.917	0.083	0.021		90		1.000				

All populations fitted the Hardy-.Weinberg equilibrium at both polymorphic loci.

The SNPs rs859208 and rs1043657 are among those analyzed in the international HAPMAP project, that has typed three populations (African Yorubans, Europeans of mixed origin and Asian Chinese and Japanese) for more than 3,000,000 SNPs (International Hap-Map Consortium 2007). The results obtained in these population are appended in table 3. The European and Yoruban Hap-Map sample overlap with those examined by us, whereas the oriental sample did not. Anyway the Hap-Map results are consistent with ours.

Apportionment of diversity

In order to quantify the amount of variation within and among continents we used the analysis of molecular variance on the compilation of two-loci genotypes. As observed with most loci, the largest quota of diversity is among individuals within populations (68.34%). We found a large among-continents quota of variation. This can be attributed to the differential distributions of frequencies at the two loci, both contributing to differentiation in this analysis. In fact, rs859208 distinguishes Africa while rs1043657 distinguishes Europe.

In order to further analyse variation within continents, we treated them separately (see table 3) Africa showed by far the highest internal heterogeneity (Fst=0.25) accounted for by allele frequencies variation at rs859208.

Table 3 results of AMOVA in 52 population samples typed for rs859208 and rs1043657

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
Among continents	4	36.688	0.02925 Va	21.90
Among populations within continents	47	24.915	0.01304 Vb	9.76
Within populations	1748	159.546	0.09127 Vc	68.34
Total	1799	221.149	0.13356	

Fixation Indices
FST : 0.31663
FSC : 0.12499
FCT : 0.21902

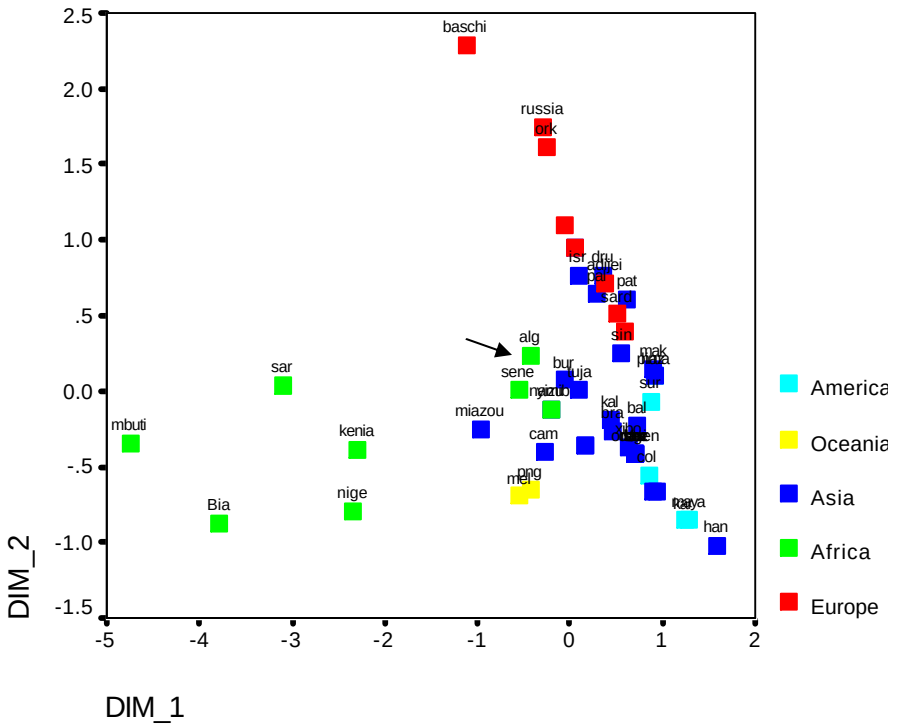
Continental heterogeneity

	d.f.among populations	d.f. within populations	Fst
Africa	7	290	0.251
Europe	7	234	0.027
Asia	28	943	0.031
America	4	205	0.004
Oceania	1	76	<0.001

We used the 52 x 52 matrix of pairwise Fst to give a representation of affinities between populations (fig. 11). In the bidimensional plot groups of populations of different continent cluster together in a pattern coerent with geography. A cluster of sub-Saharan population is clearly separated on dimension 1; Asian and American populations are on the lower right; European population are in the top center-right; Oceanian population are in the bottom center. Heterogeneity within continents can also be appreciated. The northern and western African Algerians (arrow in fig.11) and Senegalese plot afar from Bantu speakers and Pygmies. In

Europe the Basques stand out of the remaining populations, in agreement with the documented drift effect which affected this population (Wilson et al 2001; Alonso et al 2005). The rest of European populations are arranged in a north to south order from top to bottom. In Asia the position of Han agrees with the documented genetic history of this population (Wen et al 2004). Populations from extreme east Asia map at bottom right while the south east Asian Cambodians map in bottom center, close to Oceanians. Finally, native Americans map close to eastern Asian and show little heterogeneity.

Fig. 11 bidimensional plot of pairwise Fst obtained by Multi dimensional scaling



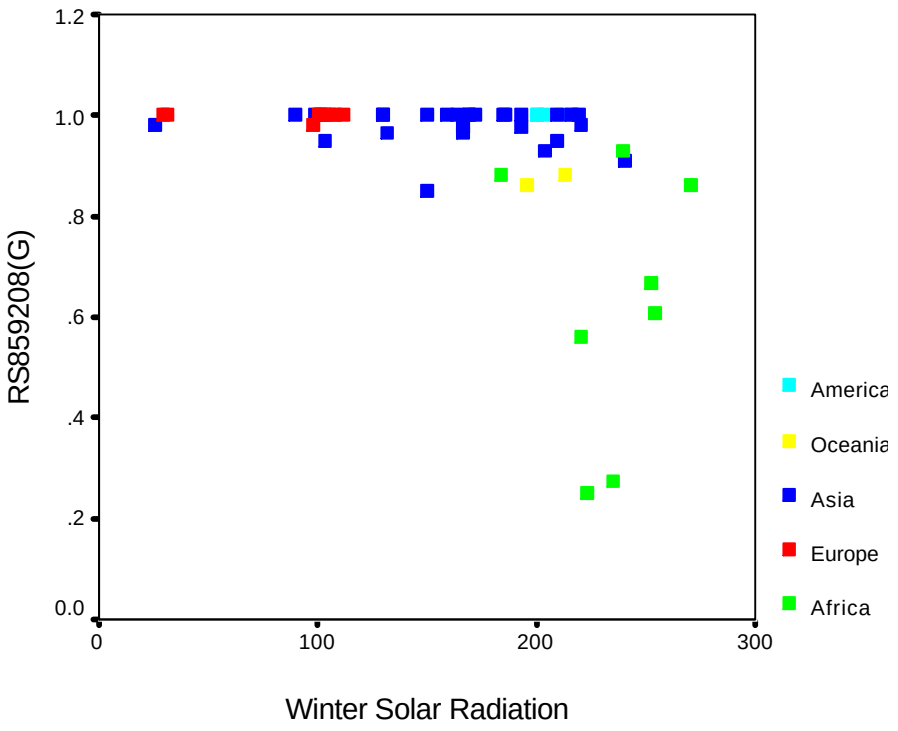
Overall, the diversity of these two loci can summarize the genetic differentiation of human populations (Li et al 2008) during their worldwide dispersal. This result can be basically attributed to the confinement of the haplotypes rs859208(C)-rs1043657(G) and rs859208(G)-rs1043657(A) to Africa and Europe, respectively. Our results show that the close genomic proximity of these two positions prevented the formation of the recombinant haplotype rs859208(C)-rs1043657(A)(see also fig. 7, left panel).

Correlates of AKR7A2: climate

The most obvious descriptors of environmental condition which imposed adaptive needs onto human population out of Africa are climatic variables. This concept has been exploited for a genome wide search for correlation, to identify SNPs candidates for genetic adaptation (Hancock et al 2008). We used the same set of variables and values for the 52 populations of the HGDP panel to search for correlation with rs859208(G) and rs1043657(G).

rs859208 showed highly significant negative correlation with a number of variables related to winter climate (average minimum and maximum temperature, average surface temperature, amount of precipitation and solar radiation; r always <-0.278). The scatterplot (fig .12) showed the predominant role of African populations in explaining such correlation.

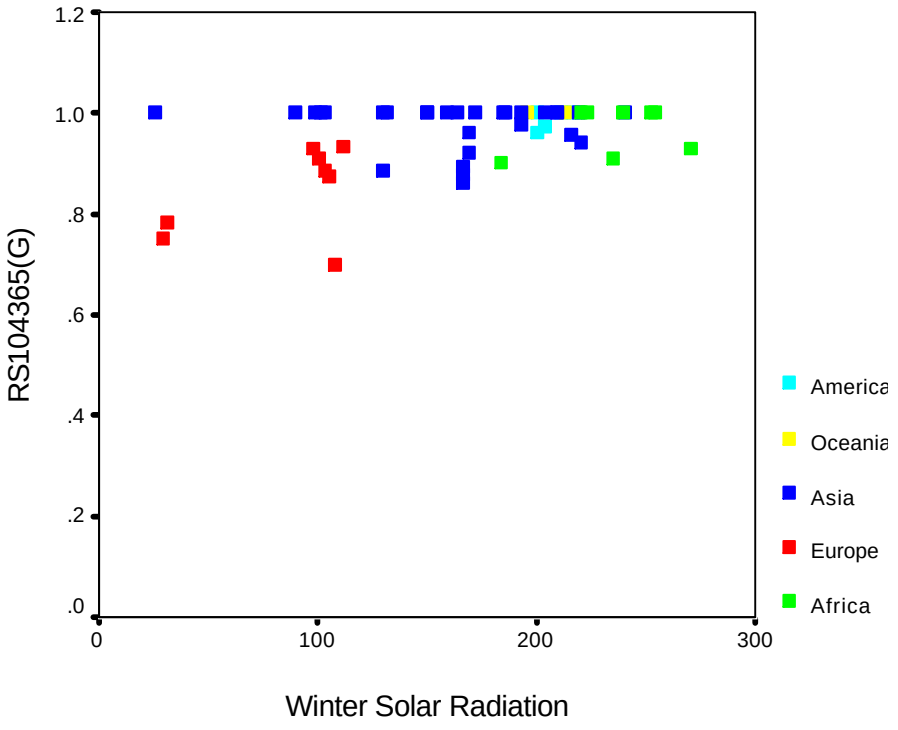
Fig. 12 plot of rs859208(G) frequency versus winter solar radiation



In fact, when African populations are removed from the analysis only correlation with winter precipitation remains significant (r=-0.397), due to the two outlier Oceanian populations (yellow dots), characterized by frequencies of 0.86-0.88 in a tropical environment.

rs104365 also showed a significant correlation with solar radiation, in this case positive (r=+0.452, p<0.001). Here the main effect is attributable to European populations. However some degree of relationship is present within continents. Within Europe, the northern Russians and Orkney have the lowest frequencies; in Africa the Algerians have the lowest frequency; within south west Asia the middle eastern populations have lower frequencies and solar radiation than the Pakistani populations.

Fig. 13 plot of rs104365(G) frequency versus winter solar radiation



Correlates of AKR7A2: genetics

Another work in this laboratory explored the intraspecific (Leone et al. 2006) variation of another gene, encoding an enzyme of the same metabolic pathway as AKR7A2, i.e. SSADH or NAD+-dependent succinic semialdehyde dehydrogenase. In this paper, two closely linked SNPs were analyzed in the HGDP panel. They showed the persistence of ancestral alleles in Africa and the near-fixation of the derived allele in Asia. We sought to analyse the correlation

of allele frequencies of the two genetic systems across populations. Fig. 14 shows a scatterplot with continental populations labelled in different colours. The overall correlation between SSADH major haplotype and rs859208(G) is positive and highly significant (r=0.509; p<0.001). Indeed, the overall correlation is the result of the outlying position of the African populations, the non-African populations being essentially non informative. On the other hand correlation between SSADH major haplotype and rs1043657 was not significant (r=0.08).

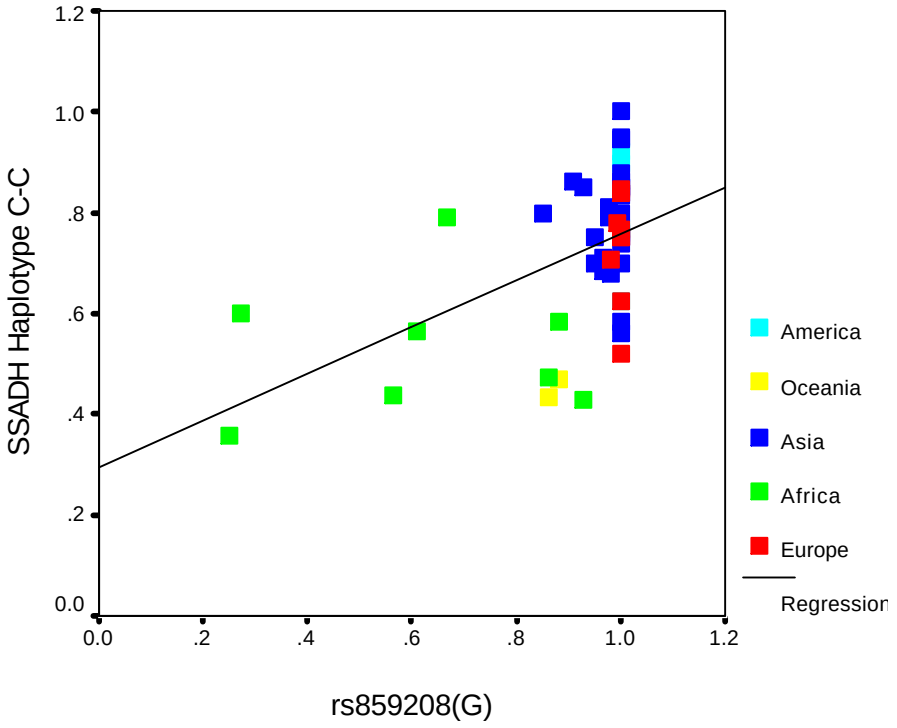


Fig. 14 analysis of correlation between SSADH major Haplotype and rs859208(G). The regression line is reported.

CONCLUSIONS

Single nucleotide polymorphisms in the AKR7A2 exon2 revealed a remarkable power in discriminating human populations. The geographical distribution of alleles at rs859208 indicates that the derived allele most likely arose and spread within Africa prior the exit out of the continent of anatomically modern humans. Conversely, the geographical pattern of distribution of the derived allele at rs104365 is compatible with a much later appearance,

possibly in the Mediterranean area and a spread which is difficult to reconstruct at the moment. Two models then compete in explaining our findings.

The first one is a purely neutral model in which populations exiting out of Africa were enriched, by chance alone, in the derived allele at rs859208. This was further purified during the subsequent migration to Asia, Europe, Oceania and Africa by a serial founder affect, as only subsets of the entire populations moved forward during this dispersal process. The phylogeographical expectations under such a model have been worked out by Currat and Excoffier (2006) and Klopftein et al. (2006). This model, which includes demographic expansions leading to increased population densities during the dispersal, generates an expected pattern of geographical variation that can be easily mistaken as the result of genetic adaptation to changing environments. In fact, the move towards eastern Asia or Europe also implies a change to more northernly and humid environments which, in principle, can also select for alleles with properties different from those required in Africa.

On the opposite extreme stands a purely selective model, in which derived alleles reached the frequencies observed today by virtue of the selective advantage they conferred by a yet unidentified mechanism. Two variants of this model can be seen. In the first variant, a new mutation generated a derived allele (in our case rs859208(G) and rs104365(A)) that is immediately favourable, as it responds to some adaptive need. In this case the derived allele starts increasing its frequency from the very beginning. Such process has been reconstructed with high likelihood for multiple alleles producing lactase persistence (and thus ability to digest dairy products) in African populations (Tishkoff et al. 2007).

Alternatively, the findings at AKR7A2 may be explained by a model of directional selection acting on standing variants, i.e. variants that segregated in the population prior to the onset of selection; these variants may have been completely neutral or slightly deleterious before they became advantageous. Due to the rapid environmental changes occurred during

human evolution, a number of investigators have postulated that selection on standing variation (rather than selection on a new beneficial allele) played a major role in human adaptations, thus affording a more rapid adaptive response to the environmental change. A variety of scenarios of directional selection on standing variation have been modelled to determine the expected signature of selection (for reviews see (Bamshad and Wooding 2003; Biswas and Akey 2006; Sabeti et al. 2006). These models may prove particularly useful to understand the pattern of variation at AKR7A2.

The current state of our data can hardly allow to distinguish between the two main models and/or between the two variants of the second one for AKR7A2. In the case of the SSADH data generated in the same laboratory, the conclusions in favour of selection as a driving agent for the worldwide pattern of allele frequencies were robust because i) an outlying positive correlation was found with a gene for which positive selection had been well documented (microcephalin) (Evans et al. 2004; Evans et al. 2005) and ii) this correlation resisted after controlling for geography, i.e. was not due to coincidental Africa-to-Asia and Africa-to-Europe trends.

Our data, however, revealed an interesting parallel between AKR7A2 and SSADH in the persistence of ancestral alleles in Africa. Thus, while conclusions on the role of natural selection in shaping this arrangement are only speculative, the two sets of data lead to the interesting perspective that African vs. non-African populations have differentiated allele repertoires in multiple steps along the same metabolic pathway. The different catalytic properties of SSADH polymorphic alleles have been demonstrated (Blasi et al. 2002), while corresponding results for AK7A2 are missing. The hypothesis that polymorphisms in the two genes co-evolved in human populations, due to their concerted functions in the GABA shunt, can then be now attacked experimentally at the functional level.

ACKNOWLEDGMENTS

I am grateful to Prof. Pina Rose and the entire staff of the genetic group of the University of Calabria for the continuous advice and support.

The useful comments by Prof. Patrizia Malaspina are gratefully acknowledged. I thank Paola Blasi for her collaboration and availability during the experimental work in the laboratories of University “Tor Vergata”. I thank Luisa, Roberta, Beatriz and Fiorenza for their friendly assistance.

A special thank to Prof. Andrea Novelletto for his kind help during the PhD course and for the interest that he transmitted me.

Finally, special thanks to my parents and my sisters who have always believed in me, encouraging to resist the most difficult moments.

To my two children for their patience when their mom was busy with this work.

REFERENCES

Alonso S, Flores C, Cabrera V, Alonso A, Martín P, Albarrán C, Izagirre N, de la Rúa C, García O.*The place of the Basques in the European Y-chromosome diversity landscape.*Eur J Hum Genet. 2005 Dec;13(12):1293-302.

Bamshad M, Wooding SP. *Signatures of natural selection in the human genome*. Nat Rev Genet. 2003 Feb;4(2):99-111.

Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, Bertorelle G, Chamberlain AT, Nunes TP, Metodiev S, Ferrand N, Erhardt G. *Gene-culture coevolution between cattle milk protein genes and human lactase genes.* Nat Genet. 2003 Dec;35(4):311-3. Epub 2003 Nov 23. Review. Erratum in: Nat Genet. 2004 Jan;36(1):106

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. *Genetic signatures of strong recent positive selection at the lactase gene.* Am J Hum Genet. 2004 Jun;74(6):1111-20. Epub 2004 Apr 26

Bianchi NO, Catanesi CI, Bailliet G, Martinez-Marignac VL, Bravi CM, Vidal-Rioja LB, Herrera RJ, López-Camelo JS. *Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations.* Am J Hum Genet. 1998 Dec;63(6):1862-71.

Biswas S, Akey JM.*Genomic insights into positive selection.*Trends Genet. 2006 Aug;22(8):437-46. Epub 2006 Jun 30.

Blasi P, Boyl PP, Ledda M, Novelletto A, Gibson KM, Jakobs C, Hogema B, Akaboshi S, Loreni F, Malaspina P. *Structure of human succinic semialdehyde dehydrogenase gene: identification of promoter region and alternatively processed isoforms.* Mol Genet Metab. 2002 Aug;76(4):348-62.

Blasi P, Palmerio F, Aiello A, Rocchi M, Malaspina P, Novelletto A. *SSADH variation in primates: intra- and interspecific data on a gene with a potential role in human cognitive functions.* J Mol Evol. 2006 Jul;63(1):54-68. Epub 2006 Jun 17.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG. *Natural selection on protein-coding genes in the human genome*. Nature. 2005 Oct 20;437(7062):1153-7. Cambridge, UK.

Cavalli-Sforza LL, Feldman MW. *The application of molecular genetic approaches to the study of human evolution.* Nat Genet. 2003 Mar;33 Suppl:266-75.

Charlesworth B, Morgan MT, Charlesworth D.*The effect of deleterious mutations on neutral molecular variation*.Genetics. 1993 Aug;134(4):1289-303.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G. *Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool.* Proc Natl Acad Sci U S A. 1998 Jul 21;95(15):9053-8.

Chimpanzee Sequencing and Analysis Consortium. *Initial sequence of the chimpanzee genome and comparison with the human genome* .Nature. 2005 Sep 1;437(7055):69-87
.

Currat M, Excoffier L, Maddison W, Otto SP, Ray N, Whitlock MC, Yeaman S. Comment on *"Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans"*. Science 2006. 313:172
.

Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, Mahowald M, Wyckoff GJ, Malcom CM, Lahn BT. *Accelerated evolution of nervous system genes in the origin of Homo sapiens.* Cell. 2004 Dec 29;119(7):1027-40.

Evans PD, Anderson JR, Vallender EJ, Choi SS, Lahn BT. *Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size*. Hum Mol Genet. 2004 Jun 1;13(11):1139-45. Epub 2004 Mar 31.

Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, Vaez-Azizi LM, Tishkoff SA, Hudson RR, Lahn BT. *Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans.* Science. 2005 Sep 9;309(5741):1717-20.

Excoffier L, Smouse PE, Quattro JM. *Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data*. Genetics. 1992 Jun;131(2):479-91.

Excoffier L. *Human demographic history: refining the recent African origin model.* Curr Opin Genet Dev. 2002 Dec;12(6):675-82.

Fay JC, Wu CI. *Hitchhiking under positive Darwinian selection*. Genetics. 2000 Jul;155(3):1405-13.

Fay JC, Wyckoff GJ, Wu CI. *Positive and negative selection on the human genome*. Genetics. 2001 Jul;158(3):1227-34.

Garrigan D, Hammer MF. *Reconstructing human origins in the genomic era*. Nat Rev Genet. 2006 Sep;7(9):669-80.

Gusmão L, Sánchez-Diz P, Calafell F, Martín P, Alonso CA, Alvarez-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. *Mutation rates at Y chromosome specific microsatellites.* Hum Mutat. 2005 Dec;26(6):520-8.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A. *Adaptations to climate in candidate genes for common metabolic disorders*. PLoS Genet. 2008 Feb;4(2):e32.

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P. *Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees*. Hum Mol Genet. 1997 May;6(5):799-803.

Hudson RR, Kreitman M, Aguadé M. *A test of neutral molecular evolution based on nucleotide data. Genetics*. 1987 May;116(1):153-9.

Hughes AL, Hughes MK, Howell CY, Nei M. *Natural selection at the class II major histocompatibility complex loci of mammals*. Philos Trans R Soc Lond B Biol Sci. 1994 Nov 29;346(1317):359-66; discussion 366-7.

Hughes AL, Nei M. *Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection.* Proc Natl Acad Sci U S A. 1989 Feb;86(3):958-62.

Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988 Sep 8;335(6186):167-70.

Jobling MA, Tyler-Smith C. *The human Y chromosome: an evolutionary marker comes of age.* Nat Rev Genet. 2003 Aug;4(8):598-612.

Karafet TM, Mendez FL, Meilerman MB et al (2008) *New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree*. Genome Res

Karafet T, de Knijff P, Wood E, Ragland J, Clark A, Hammer MF. *Different patterns of variation at the X- and Y-chromosome-linked microsatellite loci DXYS156X and DXYS156Y in human populations.* Hum Biol. 1998 Dec;70(6):979-92.

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, Redd AJ, Zegura SL, Hammer MF. *Paternal population history of East Asia: sources, patterns, and microevolutionary processes.* Am J Hum Genet. 2001 Sep;69(3):615-28. Epub 2001 Jul 30.

Kayser M, Brauer S, Stoneking M. *A genome scan to detect candidate regions influenced by local natural selection in human populations.*.Mol Biol Evol. 2003 Jun;20(6):893-900. Epub 2003 Apr 25.

Kayser M, Brauer S, Weiss G, Schiefenhövel W, Underhill PA, Stoneking M. *Independent histories of human Y chromosomes from Melanesia and Australia*. Am J Hum Genet. 2001 Jan;68(1):173-190. Epub 2000 Dec 12

Kelly VP, Sherratt PJ, Crouch DH, Hayes JD. *Novel homodimeric and heterodimeric rat gamma-hydroxybutyrate synthases that associate with the Golgi apparatus define a distinct subclass of aldo-keto reductase 7 family proteins.* Biochem J. 2002 Sep 15;366(Pt 3):847-61.

Kimura (1985) *Neutral theory of molecular evolution*. Cambridge Univ Press,

Klopfstein S, Currat M, Excoffier L.The fate of mutations surfing on the wave of a range expansion. .Mol Biol Evol. 2006 Mar;23(3):482-90. Epub 2005 Nov 9.

Knight LP, Primiano T, Groopman JD, Kensler TW, Sutter TR. *cDNA cloning, expression and activity of a second human aflatoxin B1-metabolizing member of the aldo-keto reductase superfamily, AKR7A3*. Carcinogenesis. 1999 Jul;20(7):1215-23.

Leone O, Blasi P, Palmerio F, Kozlov AI, Malaspina P, Novelletto A. *A human derived SSADH coding variant is replacing the ancestral allele shared with primates.* Ann Hum Biol. 2006 Sep-Dec;33(5-6):593-603.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh S, Feldman M, Cavalli-Sforza LL, Myers RM.*Worldwide human relationships inferred from genome-wide patterns of variation*.Science. 2008 Feb 22;319(5866):1100-4.

Livingstone FB. *Malaria and human polymorphisms*. Annu Rev Genet. 1971;5:33-64.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. *The power and promise of population genomics: from genotyping to genome typing.* Nat Rev Genet. 2003 Dec;4(12):981-94.

Lyon RC, Johnston SM, Watson DG, McGarvie G, Ellis EM. *Synthesis and catabolism of gamma-hydroxybutyrate in SH-SY5Y human neuroblastoma cells: role of the aldo-keto reductase AKR7A2.* J Biol Chem. 2007 Sep 7;282(36):25986-92. Epub 2007 Jun 25.

Malaspina P, Cruciani F, Santolamazza P, Torroni A, Pangrazio A, Akar N, Bakalli V, Brdicka R, Jaruzelska J, Kozlov A, Malyarchuk B, Mehdi SQ, Michalodimitrakis E, Varesi L, Memmi MM, Vona G, Villems R, Parik J, Romano V, Stefan M, Stenico M, Terrenato L, Novelletto A, Scozzari R. *Patterns of male-specific inter-population divergence in Europe, West Asia and North Africa.* Ann Hum Genet. 2000 Sep;64(Pt 5):395-412.

Malaspina P, Persichetti F, Novelletto A, Iodice C, Terrenato L, Wolfe J, Ferraro M, Prantera G. *The human Y chromosome shows a low level of DNA polymorphism*. Ann Hum Genet. 1990 Oct;54(Pt 4):297-305.

Kayser , Lao, Anslinger Augustin, Bargel, Elias, Heinrich, Henke, Ploski. *Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis.* Hum Genet (2005) 117: 428–443 DOI 10.1007/s00439-005-1333-9

McDonald JH, Kreitman M. *Adaptive protein evolution at the Adh locus in Drosophila*. Nature. 1991 Jun 20;351(6328):652-4.

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. *Natural selection shaped regional mtDNA variation in humans*. Proc Natl Acad Sci U S A. 2003 Jan 7;100(1):171-6. Epub 2002 Dec 30.

Neel JV. *Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?* Am J Hum Genet. 1962 Dec;14:353-62.

Pearl PL, Novotny EJ, Acosta MT et al (2003) *Succinic semialdehyde dehydrogenase deficiency in children and adults.* Ann Neurol 54 Suppl 6:S73-80

Praml C, Savelyeva L, Schwab M. *Aflatoxin B1 aldehyde reductase (AFAR) genes cluster at 1p35-1p36.1 in a region frequently altered in human tumour cells*. Oncogene. 2003 Jul 24;22(30):4765-73.

Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G. *Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms.* Genome Res. 2002 Apr;12(4):602-12.

Rosenberg NA, Nordborg M. *Genealogical trees, coalescent theory and the analysis of genetic polymorphisms.* Nat Rev Genet. 2002 May;3(5):380-90.

Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC. *Effects of purifying and adaptive selection on regional variation in human mtDNA.* Science. 2004 Jan 9;303(5655):223-6.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. *Detecting recent positive selection in the human genome from haplotype structure*. Nature. 2002 Oct 24;419(6909):832-7. Epub 2002 Oct 9.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. *Positive natural selection in the human lineage*. Science. 2006 Jun 16;312(5780):1614-20.

Schaller M, Schaffhauser M, Sans N, Wermuth B. *Cloning and expression of succinic semialdehyde reductase from human brain. Identity with aflatoxin B1 aldehyde reductase.* Eur J Biochem. 1999 Nov;265(3):1056-60.

Schneider et al (2000) ARLEQUIN v 2000: *a software for population genetics data analysis;* Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Scozzari R, Cruciani F, Malaspina P, Santolamazza P, Ciminelli BM, Torroni A, Modiano D, Wallace DC, Kidd KK, Olckers A, Moral P, Terrenato L, Akar N, Qamar R, Mansoor A, Mehdi SQ, Meloni G, Vona G, Cole DE, Cai W, Novelletto A. *Differential structuring of human populations for homologous X and Y microsatellite loci.* Am J Hum Genet. 1997 Sep;61(3):719-33.

Seielstad MT, Minch E, Cavalli-Sforza LL. *Genetic evidence for a higher female migration rate in humans.* Nat Genet. 1998 Nov;20(3):278-80.

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. *Population genetic implications from sequence variation in four Y chromosome genes.* Proc Natl Acad Sci USA. 2000 Jun 20;97(13):7354-9.

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. *The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.* Nature. 2003 Jun 19;423(6942):825-37.

Smith NG, Eyre-Walker A. *Adaptive protein evolution in Drosophila*. Nature. 2002 Feb 28;415(6875):1022-4.

Smith NG, Eyre-Walker A. *The compositional evolution of the murid genome*. J Mol Evol. 2002 Aug;55(2):197-201.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF. *Haplotype variation and linkage disequilibrium in 313 human genes*. Science. 2001 Jul 20;293(5529):489-93. Epub 2001 Jul 12. Erratum in: Science 2001 Aug 10;293(5532):104.

Stephens M, Smith NJ, Donnelly P. *A new statistical method for haplotype reconstruction from population data.* Am J Hum Genet. 2001 Apr;68(4):978-89. Epub 2001 Mar 9

Tajima F. *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism*. Genetics. 1989 Nov;123(3):585-95.

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW. *Recent common ancestry of human Y chromosomes: evidence from DNA sequence data*. Proc Natl Acad Sci 2000 USA 97:7360–7365.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P. *Convergent adaptation of human lactase persistence in Africa and Europe.* Nat Genet. 2007 Jan;39(1):31-40. Epub 2006 Dec 10.

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG. *Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance*. Science. 2001 Jul 20;293(5529):455-62. Epub 2001 Jun 21.

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

Vasiliou V, Pappa A, Estey T. *Role of human aldehyde dehydrogenases in endobiotic and xenobiotic metabolism.* Drug Metab Rev. 2004 May;36(2):279-99.

Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA. *Evidence for balancing selection from nucleotide sequence analyses of human G6PD.* Am J Hum Genet. 2002 Nov;71(5):1112-28. Epub 2002 Oct 11.

Wen B, Li H, Lu D, Song X, Zhang F, He Y, Li F, Gao Y, Mao X, Zhang L, Qian J, Tan J, Jin J, Huang W, Deka R, Su B, Chakraborty R, Jin L.*Genetic evidence supports demic diffusion of Han culture.Nature.* 2004 Sep 16;431(7006):302-5.

Wilder JA, Mobasher Z, Hammer MF. *Genetic evidence for unequal effective population sizes of human females and males.* Mol Biol Evol. 2004 Nov;21(11):2047-57. Epub 2004 Aug 18.

Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N, Goldstein DB. *Population genetic structure of variable drug response*. Nat Genet. 2001 Nov;29(3):265-9.

Wilson JF, Weiss DA, Richards M, Thomas MG, Bradman N, Goldstein DB.*Genetic evidence for different male and female roles during cultural transitions in the British Isles.*Proc Natl Acad Sci U S A. 2001 Apr 24;98(9):5078-83. Epub 2001 Apr.

Yang Z, Nielsen R. *Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.* Mol Biol Evol. 2002 Jun;19(6):908-17.

Yang Z, Nielsen R. *Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.* Mol Biol Evol. 2000 Jan;17(1):32-43.

Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L. *The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time*. Am J Hum Genet. 2004 Jan;74(1):50-61. Epub 2003 Dec 19.