

Margarida Coelho · Donata Luiselli · Giorgio Bertorelle  
Ana Isabel Lopes · Susana Seixas  
Giovanni Destro-Bisol · Jorge Rocha

## Microsatellite variation and evolution of human lactase persistence

Received: 19 January 2005 / Accepted: 8 April 2005 / Published online: 1 June 2005  
© Springer-Verlag 2005

**Abstract** The levels of haplotype diversity within the lineages defined by two single-nucleotide polymorphisms (SNPs) (–13910 C/T and –22018 G/A) associated with human lactase persistence were assessed with four fast-evolving microsatellite loci in 794 chromosomes from Portugal, Italy, Fulbe from Cameroon, São Tomé and Mozambique. Age estimates based on the intraallelic microsatellite variation indicate that the –13910\*T allele, which is more tightly associated with lactase persistence, originated in Eurasia before the Neolithic and after the emergence of modern humans outside Africa. We detected significant departures from neutrality for

the –13910\*T variant in geographically and evolutionary distant populations from southern Europe (Portuguese and Italians) and Africa (Fulbe) by using a neutrality test based on the congruence between the frequency of the allele and the levels of intraallelic variability measured by the number of mutations in adjacent microsatellites. This result supports the role of selection in the evolution of lactase persistence, ruling out possible confounding effects from recombination suppression and population history. Reevaluation of the available evidence on variation of the –13910 and –22018 loci indicates that lactase persistence probably originated from different mutations in Europe and most of Africa, even if 13910\*T is not the causal allele, suggesting that selective pressure could have promoted the convergent evolution of the trait. Our study shows that a limited number of microsatellite loci may provide sufficient resolution to reconstruct key aspects of the evolutionary history of lactase persistence, providing an alternative to approaches based on large numbers of SNPs.

**Electronic supplementary material** Supplementary material is available for this article at <http://dx.doi.org/10.1007/s00439-005-1322-z>

M. Coelho · S. Seixas · J. Rocha  
Instituto de Patologia e Imunologia Molecular da  
Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n,  
4200-465 Porto, Portugal

M. Coelho · J. Rocha (✉)  
Departamento de Zoologia Antropologia,  
Faculdade de Ciências, Universidade do Porto,  
Porto, Portugal  
E-mail: [jrocha@ipatimup.pt](mailto:jrocha@ipatimup.pt)  
Tel.: + 351-22557-0700  
Fax: + 351-22557-0799

D. Luiselli  
Dipartimento di Biologia Evoluzionistica Sperimentale,  
Università di Bologna, Bologna, Italia

G. Bertorelle  
Sezione di Biologia Evolutiva, Dipartimento di Biologia,  
Università di Ferrara, Ferrara, Italia

A. I. Lopes  
Unidade de Gastroenterologia Pediátrica,  
Hospital de Santa Maria, Lisbon, Portugal

G. Destro-Bisol  
Dipartimento di Biologia Animale e dell' Uomo,  
Università "La Sapienza", Rome, Italy

G. Destro-Bisol  
Istituto Italiano di Antropologia,  
Rome, Italy

### Introduction

The ability to digest lactose in adults is an autosomal dominant hereditary condition caused by the persistence of lactase activity in the small intestine after weaning. The frequency of lactase persistence, as evaluated by different physiological tests, varies widely in human populations and is well correlated with the distribution of dairy farming (reviewed in Swallow 2003). In the majority of populations, the ancestral mammalian developmental pattern prevails, and most people have a marked decline in lactase levels after infancy (lactase restriction), which may limit their use of large amounts of fresh milk in adulthood. In Europe, the highest frequencies of lactase persistence are observed in north-western populations, where milk-dependent cattle pastoralism was developed very early (Midgley 1992),

and there is a decrease in prevalence towards the south and east. In Africa, both north and south of the Sahara, lactase persistence is typically much more frequent among pastoralists than in neighboring non-pastoralist communities.

There are different views on the microevolutionary forces underlying the present-day distribution of lactase persistence. Several studies have proposed that the match between the geographic distribution of lactase persistence and dairy farming could be the result of the recent selective pressure associated with the added nutritional benefit of high milk consumption in populations that shifted their subsistence patterns to become crucially dependent on milk (Simoons 1970; McCracken 1971; Kretchmer 1972; Flatz 1987; Holden and Mace 1997). A quite different view is sustained by Nei and Saitou (1986), who questioned the role of selection based on the assumption that the origin of lactase persistence predated the geographical dispersion of modern humans and on the lack of a sufficiently long period of time for selection to act since the introduction of dairying. According to this interpretation, the differences in lactase persistence among human populations arose by genetic drift and preceded the major changes in subsistence patterns associated with the Neolithic. In this case, the correlation between lactase persistence and milk-based pastoralism could be either entirely fortuitous or caused by the adoption of milk drinking habits only by those populations with the ability to digest lactose (Bayless 1971; McCracken 1971; Aoki 2001).

Recently, the T allele of a C/T polymorphism in a potential regulatory site located 13,910 bp upstream the lactase gene was found to be completely associated with lactase persistence in Northern Europeans (Enattah et al. 2002). A significant, although less strong association, was also observed with a second –22,018-bp G → A mutation (Enattah et al. 2002). Besides creating a new tool for assessing lactose-digesting capability through single-nucleotide polymorphism (SNP) genotyping, these findings provided a basis for studying the major forces that shaped the distribution of lactase persistence, namely through the analysis of the levels of mutational heterogeneity and haplotype diversity associated with this trait.

In the Eurasian populations studied so far, the frequencies of the –13910\*T allele are concordant with the expectations from physiological tests (Swallow 2003). However, with the exception the Fulbe and Hausa from Cameroon, the –13910\*T allele was found to be rare in many African pastoralist communities where high frequencies of lactase persistence were previously found by using physiological tests (Mulcare et al. 2004). It is still to be demonstrated whether the observed discrepancy is caused by the fact that the allele is not causative or that lactase persistence had separate mutational origins in Africa and in Eurasia. The latter hypothesis clearly favors a prominent role of selection, since it would be unlikely that different persistence mutations might have risen in frequency and spread throughout human

dairying societies without being driven by an adaptive advantage.

Haplotype diversity studies in populations of Northern European ancestry have shown that most –13910\*T and –22018\*A alleles lie in an extended SNP-defined haplotype that was found to be unusually long for its frequency, indicating that there was not enough time for recombination to break it down (Poulter et al. 2003; Bersaglieri et al. 2004). This finding is consistent with the hypothesis that the current distribution of lactase persistence in Northern Europeans was caused by recent positive selection, but the possible confounding effects of allele-specific recombination suppression and/or population history on the extent of linkage disequilibrium need to be ruled out (Hollox 2004).

Here we present an analysis of the genetic variation and the evolutionary history of lactase persistence based on a microsatellite approach. By applying a neutrality test based on the intraallelic accumulation of mutations, which is not influenced by recombination suppression, we provide evidence of selection acting on the –13,910 kb\*T allele in four ethnically diverse populations from Europe (Portuguese, Italians, and Finnish) and Africa (Fulbe), whose heterogeneity makes the role of population history an unlikely confounding factor. Furthermore, we reevaluate the available evidence on variation of the –13910 and –22018 SNPs and conclude that lactase persistence probably originated from different mutations in Europe and most Africa even if the –13910\*T is not the causal allele. Our study shows that a battery of four microsatellite loci that can be easily typed in large samples is able to capture the information necessary to reconstruct the evolution of lactase persistence in human populations, providing an alternative to approaches based on large numbers of SNPs.

---

## Materials and methods

### Populations

DNA samples were obtained upon informed consent from Central Italy ( $n=67$  individuals; 37 from Tocco da Casauria and 30 from Rome), Northern Portugal ( $n=90$ ), the Fulbe ethnic group from Cameroon ( $n=51$ ), São Tomé Island in the Gulf of Guinea ( $n=142$ ; from different locations in the Island), and Mozambique ( $n=47$ ; from speakers of the Ronga Bantu language from Maputo).

The Fulbe sample was obtained in the province of the Extreme Nord in Cameroon, in the villages of Marua, Meme, and Mora. This population descends from nomadic herders that moved from Nigeria to the Cameroon from the eighteenth century onwards and progressively abandoned sheep farming to become settled agriculturists (Spedini et al. 1999). The samples from Mozambique and São Tomé are from populations that have neither traditions of pastoralism nor dairy practices, but provide useful information on the distri-

bution of background microsatellite haplotype variability associated with the lactase gene. Mozambique lies at the southeastern edge of the Bantu expansion and might have been a contact zone between Bantu-speaking farmers and more ancestral Khoisan (Salas et al. 2002). São Tomé started to be peopled by the end of the fifteenth century with slaves imported by Portuguese colonists from the adjacent coasts of the Gulf of Guinea and the Congo–Angola area. As a consequence of this settlement pattern this insular population has retained the high levels of genetic diversity that are generally observed in the African mainland (Tomás et al. 2002).

### SNP and microsatellite typing

Haplotype diversity was assessed through the analysis of the two SNPs associated with lactase persistence (–13910 C/T and –22018 G/A) and four linked microsatellites: D2S3010 [a (TATC)<sub>n</sub> repeat], D2S3013 [a (TA)<sub>n</sub> repeat], D2S3015 [a (CAAAA)<sub>n</sub> repeat] and D2S3016 [a (TG)<sub>n</sub> repeat] (Fig. 1).

The SNPs were typed by PCR-restriction fragment length polymorphism (PCR-RFLP) methods. The –13910 C/T polymorphism was amplified within a 125-bp fragment with primers 5'-GCAGGGCTCAAA-GAACAATC-3' (forward) and 5'-TGTACTAGTAGG-CCTCTGCGCT-3' (reverse). The –13910\*T allele introduces a *Bsm*FI restriction site that originates digestion product sizes of 80 and 45 bp. The –22018 G/A locus was amplified within a 271-bp product with primers 5'-CTCAGTGATCCTCCCACCTC-3' and 5'-CCCCTACCCTATCAGTAAAGGC-3'. Digestion with *Hin*6I generates 196- and 75-bp fragments in the presence of the –22018\*G allele. PCR reactions contained 0.5 μM of each primer, 0.2 mM of each deoxynucleotide triphosphate (dNTP), 10 mM Tris–HCl (pH 8.8), 50 mM KCl, 0.08% Nonidet, 1.5 mM MgCl<sub>2</sub> (1.0 mM for the –22018 G/A locus) and 1 U *Taq* polymerase. Samples were denatured for 5 min at 94°C, followed by 35 cycles of 94°C for 1 min, 58°C for 1 min, and 72°C for 1 min,

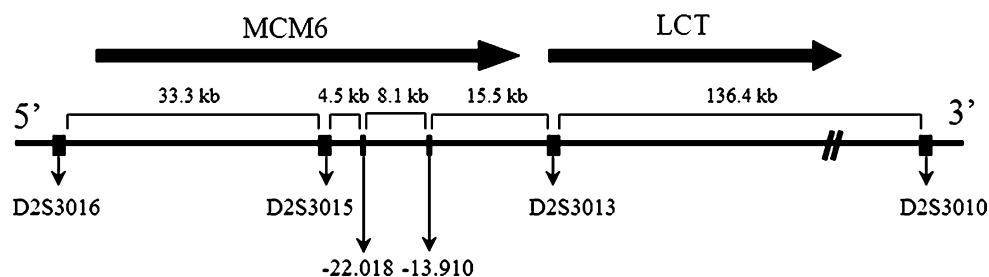
followed by a 20-min extension at 72°C. Digestions (1 U/μl) were performed at 50°C for 1 h and DNA fragments were visualized by silver staining after non-denaturing electrophoresis separation in 9% polyacrylamide gels.

Microsatellites were typed by PCR amplification in two duplex reactions followed by separation of amplification products in an ABI 310 DNA sequencer. Fragment analysis and weight determination were performed with the GeneScan software. The first duplex reaction included the primers for D2S3013 (5'-GAGA-ATATAGTCATAAACTATGTT-3' and 5'-ATT-TTGATTATATATGCTTTCTTG-3', labeled with FAM fluorescence) and D2S3015 (5'-CCTGTAGTCC-CAGCTAATTTTC-3' and 5'-CAGAGAAGTTTTGTT-TGTGGA-3', labeled with TET fluorescence) at 0.5 and 0.075 μM concentrations, respectively. The second duplex reaction included the primers for D2S3010 (5'-TTAGGCCTCTCTTCGAATGAT-3' and 5'-GAT-TTAGGTGGAGACACAC-3', labeled with FAM fluorescence) and D2S3016 (5'-GAGAAAATTAGGT-GTGAACCA-3' and 5'-CCCTTTAGCTGCCTGA-ACTG-3', labeled with TET fluorescence) at 0.5 and 0.075 μM concentrations, respectively. All other reagents were as above. In both duplex reactions, samples were denatured for 5 min at 94°C, followed by 35 cycles of 94°C for 1 min, 55°C for 1 min, and 72°C for 1 min, followed by a 20-min extension at 72°C.

### Haplotype determination

Haplotypes of sampled individuals were reconstructed from the combined genotype data for all the populations by the statistical inference method implemented in the software package PHASE, version 2.0.2 (Stephens et al. 2001; Stephens and Donnelly 2003), which provides the probabilities of the most likely pairs of haplotypes for each individual. Haplotypes were inferred with two alternative approaches. In the first approach we determined three-locus haplotypes consisting of the two –13910 and –22018 SNPs and each microsatellite marker. In the second approach we inferred the full six-locus haplotypes, combining the two SNPs and the four microsatellites. Individual haplotype phases were assigned by choosing the most probable haplotype pair that was compatible with the individual multi-locus genotypes. The haplotype frequencies were calculated by direct counting after resolution of each individual haplotype phase.

**Fig. 1** Schematic representation of the genetic interval including the lactase locus (*LCT*) and the neighbor gene for the human homologue of a yeast gene involved in the cell cycle (*MCM6*), with the relative locations of the two SNPs (–13.910 kb and –22.018 kb) and the four microsatellites (D2S3010, D2S3013, D2S3015, D2S3016) used to characterize the haplotype diversity associated with the lactase restriction/persistence polymorphism. Distances are as in BAC clone RP11-34L23 (GenBank accession no. ACO118937) 75x23mm (300x300 DPI)



## Age estimates

To estimate the time to the most recent common ancestor (TMRCA) of the  $-13910^*T$  allele associated with lactase persistence, we used two different methods based on the intraallelic accumulation of microsatellite diversity, assuming a stepwise mutation model and using a 25-year generation time.

In the first method, an unbiased estimator of the TMRCA was calculated, assuming no recombination, by the average squared difference in repeat number between each sampled  $-13910^*T$  haplotype and the root haplotype (Stumpf and Goldstein 2001). The root of the  $-13910^*T$  clade was obtained by combining together the modal allele lengths at each microsatellite locus in the pooled sample from all the populations. The TMRCA central estimates and confidence intervals were calculated using the program Ytime (Behar et al. 2003).

The second method is based on the simulation of the overtime decay in the frequency of the allele originally associated with  $-13910^*T$  in each microsatellite locus (Seixas et al. 2001). Unlike the previous method, this approach allows for recombination to be taken into account. The modal allele length at each microsatellite locus in the pooled sample was considered to be the ancestral and the combined TMRCA was calculated as the weighted average of the single locus estimates, with the weight of each microsatellite locus determined by the sum of its corresponding mutation and recombination rates. Recombination rates ( $r$ ) were calculated using the general relation  $1\text{ cM}=1\text{ Mb}$ , according to the approximate estimates provided by Kong et al. (2002) for the region encompassing the four microsatellite loci. Confidence intervals were calculated assuming a rapid population growth according to Goldstein et al. (1999).

For each age estimation method we used two sets of microsatellite mutation rates ( $\mu$ ). The first set was derived indirectly from the parameter  $\theta=4Ne\mu$  assuming mutation-drift equilibrium and using the unbiased  $\theta$  estimator proposed by Xu and Fu (2004), based on the sample homozygosity under the single-step stepwise mutation model. We assumed  $Ne=10,000$  (Takahata 1993) and estimated homozygosities from the microsatellite allele frequency distributions in São Tomé, which are less likely to have been distorted by a possible increase in the frequency of tolerance-associated chromosomes due to selection (see below). The second set of mutation rates was derived from the average 0.001 value obtained from observed mutations in pedigrees (Weber and Wong 1993). Locus specific mutation rates were calculated by apportioning this average according to the ratios of the locus-specific estimates calculated by the indirect approach.

## Neutrality tests

To assess the role of natural selection in shaping the distribution of the  $-13910^*T$  allele, we used the test

developed by Slatkin and Bertorelle (2001), which evaluates whether the observed frequency of an allele is consistent with its levels of variability under a given demographic pattern, assuming neutrality. We used the test modality that measures the intraallelic variability by the minimum number of mutations ( $S_0$ ) observed at linked microsatellite marker loci (Slatkin and Bertorelle 2001; Slatkin 2002).

The tests were performed by considering the simultaneous combination of all four microsatellites with the  $-13910^*T$  allele. The minimum number of mutations necessary to generate the observed haplotypes ( $S_0$ ) was inferred by using median-joining networks (Bandelt et al. 1999) calculated with the program NETWORK 4.0.0.0 (<http://www.fluxus-engineering.com>). All tests were performed under a number of different demographic models (see below) with the two sets of mutation rates used for calculating the TMRCA of the  $-13910^*T$  allele.

---

## Results

### Haplotype diversity

The frequencies of the core haplotypes defined by the  $-13910\text{ C/T}$  and  $-22018\text{ G/A}$  SNPs, and the expected prevalence of lactase persistence in different populations are shown in Table 1. Estimates from northern Portugal, Italy and the Fulbe are within the frequency ranges previously reported on the basis of physiological tests (Flatz 1987; Swallow 2003). The estimates from São Tomé and Mozambique are within the range observed for the majority of African non-pastoralist populations (Flatz 1987; Swallow 2003). It is likely that the occurrence of the  $-13910^*T$  allele in these two populations is due to recent admixture with Europeans. In São Tomé, for example, the frequency of the  $-13910^*T$  allele is very close to that expected from a previously calculated 11% level of admixture with the Portuguese colonists (Tomás et al. 2002).

The C–A haplotype is rare in all samples. Since, as previously shown (Poulter et al. 2003; Swallow 2003), the  $-13910$  and  $-22018$  polymorphisms were originated according to a C–G  $\rightarrow$  C–A  $\rightarrow$  T–A phylogenetic sequence, the low frequency of this intermediate haplotype indicates that the  $-22018\text{ G} \rightarrow \text{A}$  mutation might have occurred only shortly before the  $-13910\text{ C} \rightarrow \text{T}$  mutation. It is the occasional occurrence of C–A chromosomes that may lead to the wrong identification of lactase persistence on the basis of  $-22018$  genotyping.

Microsatellite allele frequency distributions within the common C–G and T–A  $-13910/-22018$  SNP core haplotypes in a pooled sample combining the data from all populations are shown in Fig. 2. Equivalent distributions for each sample are shown in Fig. S1 of the Electronic supplementary material (ESM). The data presented were retrieved from an inferred distribution of

**Table 1** Frequencies of the haplotypes defined by the –13910 and –22018 SNPs and predicted prevalence of lactase persistence in the different populations

Haplotype <sup>a</sup>		Populations				
–13910	–22018	Portugal (n=90)	Italy <sup>b</sup> (n=67)	Fulbe (n=51)	São Tomé (n=142)	Mozambique (n=47)
C	G	0.62	0.87	0.79	0.94	0.99
C	A	0.01	–	–	0.02	–
T	A	0.37	0.13	0.21	0.04	0.01
Predicted frequency of lactase persistence <sup>c</sup>		0.62	0.24	0.38	0.08	0.02

<sup>a</sup>The haplotype frequencies were retrieved from the inferred distribution of six locus SNP/microsatellite haplotypes presented in Table S1 of the Electronic supplementary material

<sup>b</sup>Samples from Tocco da Causaria and from Rome were pooled since they are not significantly different ( $P=0.61$ ), using the exact

test of population differentiation of Raymond and Rousset (1995) implemented in the Arlequin 2.1 software (Schneider et al. 2000)

<sup>c</sup>Frequency of –13910 CT + TT genotypes assuming Hardy–Weinberg equilibrium.

six-locus full haplotypes combining the two –13910 and –22018 SNPs, and the four microsatellite markers, which had a 70% proportion of phase callings with confidence values  $\geq 75\%$  (Table S1, ESM). An alternative inference approach based on the determination of three-locus haplotypes consisting of the two SNPs and each microsatellite yielded higher fractions of haplotypes with phase calling probabilities  $\geq 75\%$ , ranging from 95% for locus D2S3010 to 99% for loci D2S3015 and D2S3016 (data not shown). However, we found no significant differences between the microsatellite allele frequency distributions within the core SNP haplotypes obtained by the two approaches, using an exact test of population differentiation (Raymond and Rousset 1995). Moreover, no significant differences were found between the confidence of phase callings involving C–G and T–A SNP core haplotypes in each of the two approaches.

A clear reduction in microsatellite variation was found within the T–A haplotype (Fig. 2). This decrease in the –13910\*T intraallelic diversity is not uniform across all microsatellite markers and the higher variability accumulated in D2S3010 and D2S3013 suggests that these loci have higher mutation rates than D2S3015 and D2S3016 (Fig. 2). This is also confirmed by the decreased heterozygosities observed for the D2S3015 and D2S3016 markers among C–G haplotypes (Fig. 2).

A median-joining network relating the compound SNP-microsatellite haplotypes in the pooled sampled is shown in Fig. 3. The network has two main branches that reflect the bimodality of the D2S3013 microsatellite allele frequency distributions within C–G core haplotypes (Figs. 2, 3). In contrast with the high variability associated with C–G chromosomes, T–A haplotypes are tightly clustered within one of the two main branches irrespectively of their geographic location, as expected from a unique, relatively recent origin. Within the T–A clade, the microsatellite configuration 10–21–4–2 (D2S3010–D2S3013–D2S3015–D2S3016) is found in all populations, except Mozambique, and is likely to

represent the ancestral chromosome since it is the most frequent haplotype and combines the modal allele from each individual locus (Fig. 3, inset; Table S1, ESM). Alone it represents 25% of all sampled T–A chromosomes, 52% together with its four one-step neighbors (11–21–4–2; 10–22–4–2, 10–20–4–2 and 9–21–4–2).

An additional feature of the microsatellite allele frequency distributions is the apparent lack of recombinant T–A haplotypes within the 61.4-kb region encompassing the D2S3013, D2S3015 and D2S3016 loci (Fig. 1). This is indicated by the complete absence of diversity in D2S3015 and D2S3016 and by the observation of a clear unimodal distribution at the D2S3013 locus, which suggests the occurrence of a stepwise accumulation of mutations in an ancestral T–A haplotype carrying the D2S3013\*21 allele (Fig. 2). If recombination had played a major role in the generation of D2S3013 diversity, the striking bimodality observed within C–G haplotypes would be at least partially reflected among the T–A chromosomes and these would not cluster just in one side of the haplotype network (Figs. 2, 3). Due to a less clear difference between the shape of the D2S3010 microsatellite allele frequency distributions among C–G and T–A haplotypes, it is more difficult to evaluate the role of recombination in the regeneration of diversity in this locus. Taken together, these observations highlight the usefulness of using faster evolving markers to subtype haplotypes that could be otherwise homogeneous if defined only by SNPs.

#### TMRCAs of the –13910\*T allele

The TMRCAs of the –13910\*T allele calculated in different samples are presented in Table 2. Calculations for the Finnish sample were performed with data taken from Enattah et al. (2002) and do not include locus D2S3010.

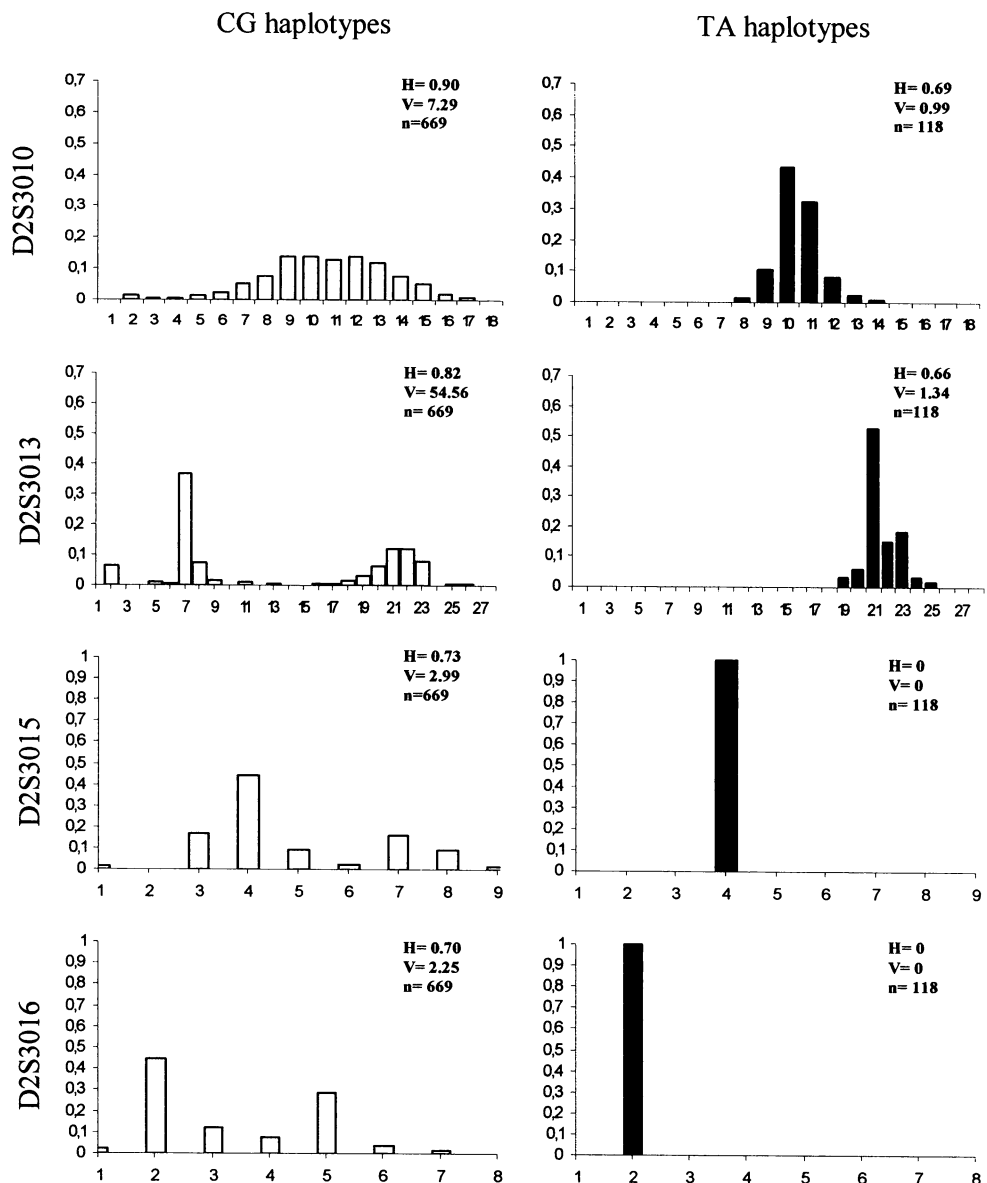
Since age calculations can be typically affected by the misestimation of mutation and recombination rates

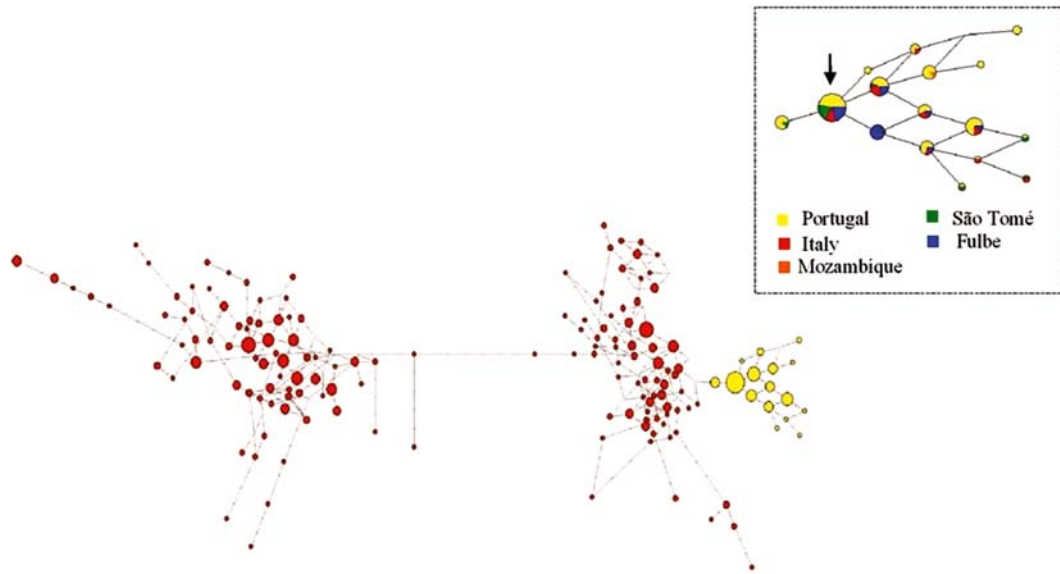
(Reich and Goldstein 1999), we considered different combinations of these parameters to assess the sensitivity of estimates to variation in their values. In the pooled sample, estimates that do not take recombination into account ( $m_1$  and  $m_2$ ) are within 45,000–30,000 or 17,500–11,750 year ranges, depending on the use of indirect or direct estimates of microsatellite mutation rates, respectively. In general the average square distance method ( $\Delta$ ) leads to higher age estimates than calculations based on the decrease in frequency of the modal microsatellite allele ( $p$ ). This discrepancy is particularly noticeable in the sample from São Tomé and may be due to the presence of low frequency microsatellite alleles that are several mutational steps away from the modal haplotype and which may not be taken into account in the “ $p$  method”.

If we assume that both recombination and mutation contribute to the intraallelic diversity, the observed haplotype homogeneity and lack of T–A recombinant chromosomes can be only explained by a more recent TMRCA of the –13910\*T allele. Accordingly, calculations based on the decrease in frequency of the modal microsatellite allele that do not assume recombination suppression lead to the lowest TMRCA estimates: 12,300 and 7,450 years in the pooled sample, under assumption sets  $m_3$  and  $m_4$ , respectively (Table 2).

Age estimates in the Fulbe and Finnish samples were found to be lower than in Portugal and Italy. These differences reflect the interpopulation variation in the levels of intraallelic diversity that may be caused by the combined effects of sampling variance and specific demographic histories of each population, like differ-

**Fig. 2** Microsatellite allele frequency distributions within C-G and T-A –13910/–22018 SNP haplotypes in a pooled sample from São Tomé, Mozambique, Portugal, Italy and Cameroonian Fulbe. The estimated sizes of allele 1 in each microsatellite are: 188 bp for D2S3010; 133 bp for D2S3013; 190 bp for D2S3015 and 148 bp for D2S3016.  $H$  heterozygosity;  $V$  variance in repeat number;  $n$  number of chromosomes 148×177mm (300×300 DPI)





**Fig. 3** Median-joining network (Bandelt et al. 1999) representing the compound SNP-microsatellite haplotype variation in a pooled sample from São Tomé, Mozambique, Portugal, Italy and Cameroonian Fulbe. C-G haplotypes are shown in red and T-A haplotypes are shown in yellow. The distribution of haplotype variation within T-A chromosomes is shown in the *inset*. Haplotypes are represented by *circles*, with areas proportional to the number of individuals harboring the haplotype. The putative ancestral 10-21-4-2 (D2S3010-D2S3013-D2S3015-D2S3016) haplotype is indicated with an *arrow*. Networks were calculated with the program NETWORK 4.0.0.0., using the same weight for SNP and microsatellite loci and the ‘frequency >1’ option, which selects only the haplotypes that occur more than once in the data set 350×192mm (72×72 DPI)

ences in the long term population size or in levels of drift during the dispersion of the trait.

#### Neutrality tests

Table 3 presents the results of the neutrality tests for the  $-13910^*T$  allele in different samples using the method of Slatkin and Bertorelle (2001). The data consist of the full haplotypes combining all four microsatellite markers linked to the  $-13910^*T$  allele. For illustrative purposes we present the outcomes obtained with different combinations of two global demographic models (D1 and D2) and the two sets of microsatellite mutation rates used for age calculation (see Table 2). The first demographic model (D1) is based on the analysis of Pritchard et al. (1999) and assumes a constant exponential growth rate of 0.008 starting 900 generations ago from an initial population of  $10^5$ . The second model (D2) is a variation of the scenarios simulated by Kruglyak (1999) and assumes that the effective population size increased exponentially from  $10^4$  to  $5 \times 10^9$ , also starting 900 generations ago. A smaller long term population size and lower genetic diversity is expected under demographic model D1.

Neutrality is rejected at the 0.001 level for the Portuguese, Finnish and Fulbe samples under all sets of assumptions (Table 3). In the Italian sample, neutrality cannot be rejected at the same significance level under the most conservative assumption, which combines demographic model D1 with the set of lower mutation rates ( $m_1$ ), decreasing the expected levels of intraallelic diversity under neutrality. In São Tomé, where a much lower frequency of the  $-13910^*T$  allele is found, neutrality is rejected only with the assumptions involving the set of higher mutation rates ( $m_2$ ).

Similar patterns and conclusions were obtained with a variety of other reported demographic models, differing in the rate of exponential growth, time of onset of population growth, and effective population sizes before expansion (Rogers and Harpending 1992; Marjoram and Donnelly 1994; Wall and Przeworski 2000; Slatkin and Bertorelle 2001; Pluzhnikov et al. 2002) (results not shown).

#### Discussion

The distribution of the ability to digest lactose in human populations is generally claimed to be an example of genetic adaptation to recent modifications in human dietary habits. However, the support of population genetics for a causative link between dairy farming and lactase persistence has been mostly based on geographic correlations and considerable controversy still exists about the relative roles that selection and population history might have played in the origin, evolution, and spread of this trait. As a contribution to the understanding of this topical issue, we have characterized the patterns of haplotype variation within lineages defined by SNPs  $-13910$  C/T and  $-22018$  G/A, using four microsatellite loci encompassing 198 kb around the lactase gene (*LCT*) in 794 chromosomes from five ethnically diverse populations with different genetic backgrounds

**Table 2** Estimates of the time (years) to the most recent common ancestral of the  $-13910^*T$  allele

Population	Age estimation method					
	Average square distance, $\Delta$		Decrease in frequency of modal allele, $p$			
	$m_1^a$	$m_2^b$	$m_1$	$m_2$	$m_3^c$	$m_4^d$
Portugal ( $n_i = 66$ ) <sup>e</sup>	48,370 (9,910–127,870) <sup>f</sup>	18,930 (3,870–53,620)	38,940 (23,690–75,500)	15,250 (11,690–39,440)	15,560 (10,000–28,440)	9,370 (5,940–17,190)
Italy ( $n_i = 17$ )	52,220 (10,990–142,000)	20,440 (4,310–57,000)	34,625 (13,440–140,000)	13,560 (5,250–54,440)	13,560 (5,750–42,190)	8,315 (3,440–27,690)
Finland <sup>g</sup> ( $n_i = 33$ )	23,640 (0–88,125)	9,250 (0–34,000)	20,750 (9,625–39,875)	8,125 (3,750–15,625)	nd <sup>h</sup>	nd <sup>h</sup>
Fulbe ( $n_i = 21$ )	20,025 (1,475–60,075)	7,840 (575–26,125)	23,815 (9,315–54,130)	9,310 (3,625–28,250)	10,125 (4,000–26,250)	6,060 (2,440–16,810)
São Tomé ( $n_i = 13$ )	46,750 (9,625–131,810)	18,290 (3,750–54,440)	17,560 (3,940–51,690)	6,875 (1,560–20,250)	7,375 (1,750–19,000)	4,400 (1,000–11,940)
Pooled <sup>i</sup> ( $n_i = 117$ )	44,610 (10,040–110,000)	17,460 (4,125–48,300)	30,000 (21,125–43,690)	11,750 (8,250–17,125)	12,300 (8,940–17,125)	7,440 (5,375–10,440)

<sup>a</sup>Assuming suppression of recombination and microsatellite indirect estimation of mutation rates:  $\mu_1(\text{D2S3010}) = 0.0009$ ;  $\mu_2(\text{D2S3013}) = 0.0005$ ;  $\mu_3(\text{D2S3015}) = 0.000095$ ;  $\mu_4(\text{D2S3016}) = 0.00011$

<sup>b</sup>Assuming suppression of recombination and microsatellite mutation rates calculated from a 0.001 direct average estimate:  $\mu_1(\text{D2S3010}) = 0.0023$ ;  $\mu_2(\text{D2S3013}) = 0.0013$ ;  $\mu_3(\text{D2S3015}) = 0.0002$ ;  $\mu_4(\text{D2S3016}) = 0.0003$

<sup>c</sup>Mutation rates as in  $m_1$  and assuming the following recombination rates between the  $-13910$  site and each microsatellite locus:  $r_1(\text{D2S3010}) = 0.0015$ ;  $r_2(\text{D2S3013}) = 0.00016$ ;  $r_3(\text{D2S3015}) = 0.00013$ ;  $r_4(\text{D2S3016}) = 0.00046$

<sup>d</sup>Mutation rates as in  $m_2$  and recombination rates as in  $m_3$

<sup>e</sup> $n_i$  represents the number of chromosomes bearing the  $-13910^*T$  allele

<sup>f</sup>95% confidence intervals are given *in parentheses*; confidence intervals for the  $\Delta$  method were calculated assuming no population growth

<sup>g</sup>Based on the data from Enattah et al. (2002), not including locus D2S3010

<sup>h</sup>Not done, due to unavailable distribution of microsatellite allele frequencies in the general population

<sup>i</sup>Excluding Finland

and subsistence patterns. Our study infers the intraallelic genealogy of lactase persistence based on the use of mutations in fast evolving markers, yields new results concerning the role of selection in the evolution of this trait in populations outside northern Europe and presents a phylogeographic interpretation suggesting that lactase persistence most probably arose from different mutations in Europe and most of Africa.

The assessment of intraallelic diversity indicates that, even assuming recombination suppression, the TMRCA of the  $-13910^*T$  variant, which is more tightly associated with lactase persistence, is unlikely to be much older than  $\sim 45,000$  years, although it may be as recent as 12,500–7,500 years if more realistic sets of assumptions that take recombination into account are applied. Since these estimates refer only to the age since intraallelic diversity began to accumulate due to an increase in frequency or population expansion, the actual age of the  $-13910 C \rightarrow T$  mutation may be older than inferred by its TMRCA (Slatkin and Rannala 2000). However, when the TMRCA estimates are taken together with the observation of low frequencies of  $-13910^*T$  in most sub-Saharan African populations (Mulcare et al. 2004), it is reasonable to conclude that the mutation originated in Eurasia before the Neolithic and after the emergence of modern humans outside Africa 100,000–50,000 years ago (Klein 2000; Relethford 2001).

Our analysis of neutrality, using a wide range of demographic models, provides strong support to the

notion that lactase persistence underwent a rapid increase in frequency, due to a selective advantage. Given the low probability values for finding the observed intraallelic diversities under neutrality (Table 3), this conclusion is unlikely to have been affected by ascertainment bias and seems to be sufficiently robust to remain unaltered by further corrections for multiple tests.

The major implications of these results lie in the assumptions underlying the neutrality test and in the composition of the population dataset. Differently from a previous assessment of selection in the *LCT* gene based on the observation of an extended haplotype in northern European-derived populations with a panel of 100 SNPs covering 3.2 Mb (Bersaglieri et al. 2004), our approach does not rely upon recombination and measures intraallelic diversity by the number of mutations in fast evolving linked microsatellites. Therefore, our test is exempt from the possible confounding effects of allele-specific recombination rates, which may produce unusually long *LCT* haplotypes at high frequency, even in the absence of selection (Hollox 2004). The robustness of our conclusions is further strengthened by the fact that evidence for selection is not confined to a single, relatively homogeneous, northern-European population, but can be found in samples from southern Europe (Portugal and Central Italy) and in the Cameroonian Fulbe, which lie in the periphery of the distribution of the  $-13910^*T$  allele and are separated from the major regions of high frequency of lactase persistence in Europe by a belt of agricultural northern-African popula-



a further search for candidate mutations could then be restricted to the selected subhaplotypes.

**Acknowledgements** We thank Luís Pedro Resende and Cinzia Battaglia for assistance in typing the Portuguese and Italian samples, respectively. We are also grateful to Gabriella Spedini for the Fulbe DNA samples and to António Prista and Silvio Saranga for the Mozambique samples. We thank Eduardo Tarazona-Santos and Nuno Ferrand for comments on the manuscript. This research was supported by the Sociedade Portuguesa de Gastroenterologia and by the Fundação para a Ciência e a Tecnologia (grants POCTI/42510/ANT/2001 and POCTI/BIA-BDE/56654/2004). D.L. and G.D.B. were supported by the M.I.U.R. (grant numbers 2003054059 and 2005058414).

## References

- Aoki K (2001) Theoretical and empirical aspects of gene-culture coevolution. *Theor Popul Biol* 59:253–261
- Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16:37–48
- Bayless TM (1971) Junior, why didn't you drink your milk?. *Gastroenterology* 60:479–480
- Behar DM, Thomas MG, Skorecki K, Hammer MF, Bulygina E, Rosengarten D, Jones AL, Held K, Moses V, Goldstein D, Bradman N, Weale ME (2003) Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. *Am J Hum Genet* 73:768–779
- Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE and Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 74:1111–1120
- Enattah NS, Sahi T, Savilahti E, Terwilliger JS, Peltonen L, Järvelä I (2002) Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 30:233–237
- Flatz G (1987) Genetics of lactose digestion in humans. *Adv Hum Genet* 16:1–77
- Goldstein DB, Reich DE, Bradman N, Usher S, Seligsohn U, Peretz H (1999) Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. *Am J Hum Genet* 64:1071–1075
- Holden C, Mace R (1997) Phylogenetic analysis of the evolution of lactose digestion in adults. *Hum Biol* 69:605–628
- Hollox E (2004) Genetics of lactase persistence—fresh lessons in the history of milk drinking. *Eur J Hum Genet* 13:267–269
- Klein RG (2000) Archeology and the evolution of human behavior. *Evol Anthropol* 9:17–36
- Kong A, Gudbjartsson DF, Sainz J, Jonsson GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. *Nat Genet* 31:241–247
- Kretchmer N (1972) Lactose and lactase. *Sci Am* 227:70–78
- Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 22:139–144
- Marjoram P, Donnelly P (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* 136:673–683
- McCracken RD (1971) Lactase deficiency: an example of dietary evolution. *Curr Anthropol* 12:479–517
- Middleton MS (1992) TRB culture: the first farmers of the North European plain. Edinburgh University Press, Edinburgh
- Mulcare CA, Weale ME, Jones AL, Connell B, Zeitlyn D, Tarekegn A, Swallow DM, Bradman N, Thomas MG (2004) The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (*LCT*) (C–13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am J Hum Genet* 74:1102–1110
- Nei M, Saitou N (1986) Genetic relationship of human populations and ethnic differences in relation to drugs and food. In: Kalow W, Goedde HW, Agarwal DP (eds) Ethnic differences in reactions to drugs and other xenobiotics. Alan L Riss, New York, pp 21–37
- Pluzhnikov A, Di Rienzo A, Hudson R (2002) Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* 161:1209–1218
- Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarnier M, Korpela R, Swallow DM (2003) The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Ann Hum Genet* 67:298–311
- Pritchard JK, Seielstad MT, Pérez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* 16:1791–1798
- Raymond M, Rousset F (1995) An exact test for population differentiation. *Evolution* 49:1280–1281
- Reich DE, Goldstein DB (1999) Estimating the age of mutations using the variation at linked markers. In: Goldstein DB, Schlötterer C (eds) Microsatellites: evolution and applications. Oxford University Press, Oxford, pp 129–138
- Relethford JH (2001) Genetics and the search for modern human origins. Wiley, New York
- Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* 9:552–569
- Salas A, Richards M, De La Fe T, Lareu MV, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo A (2002) The making of the African mtDNA landscape. *Am J Hum Genet* 71:1082–1111
- Schneider S, Roessli D, Excoffier L (2000) Arlequin, ver. 2.000: a software for population genetics data analysis. University of Geneva, Geneva
- Seixas S, Garcia O, Trovoada MJ, Santos MT, Amorim A, Rocha J (2001) Patterns of haplotype diversity within the serpin gene cluster at 14q32.1: insights into the natural history of the alpha<sub>1</sub>-antitrypsin polymorphism. *Hum Genet* 108:20–30
- Simoons FJ (1970) Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. II. A culture historical hypothesis. *Am J Dig Dis* 15:695–710
- Slatkin M (2002) The age of alleles. In: Slatkin M, Veuille M (eds) Modern developments in theoretical population genetics: the legacy of Gustave Malécot. Oxford University Press, New York, pp 233–260
- Slatkin M, Bertorelle G (2001) The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* 158:865–874
- Slatkin M, Rannala B (2000) Estimating allele age. *Annu Rev Genomics Hum Genet* 1:225–249
- Spedini G, Destro-Bisol G, Mondovi S, Kaptué L, Taglioli L, Paoli G (1999) The peopling of sub-Saharan Africa: the case study of Cameroon. *Am J Phys Anthropol* 110:143–162
- Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73:1162–1169
- Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978–989
- Stumpf MPH, Goldstein DB (2001) Genealogical and evolutionary inference with the human Y chromosome. *Science* 291:1738–1742
- Swallow DM (2003) Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet* 37:197–219
- Takahata N (1993) Allelic genealogy and human evolution. *Mol Biol Evol* 10:2–22
- Tomás G, Seco L, Seixas S, Faustino P, Lavinha J, Rocha J (2002) The peopling of São Tomé (Gulf of Guinea): origins of slave settlers and admixture with the Portuguese. *Hum Biol* 74:397–411

- Wall JD, Przeworski M (2000) When did the human population size start increasing?. *Genetics* 155:1865–1874
- Weber JL, Wong C (1993) Mutation of human short tandem repeats. *Hum Mol Genet* 2:1123–1128
- Xu H, Fu Y (2004) Estimating effective population size or mutation rate with microsatellites. *Genetics* 166:555–563